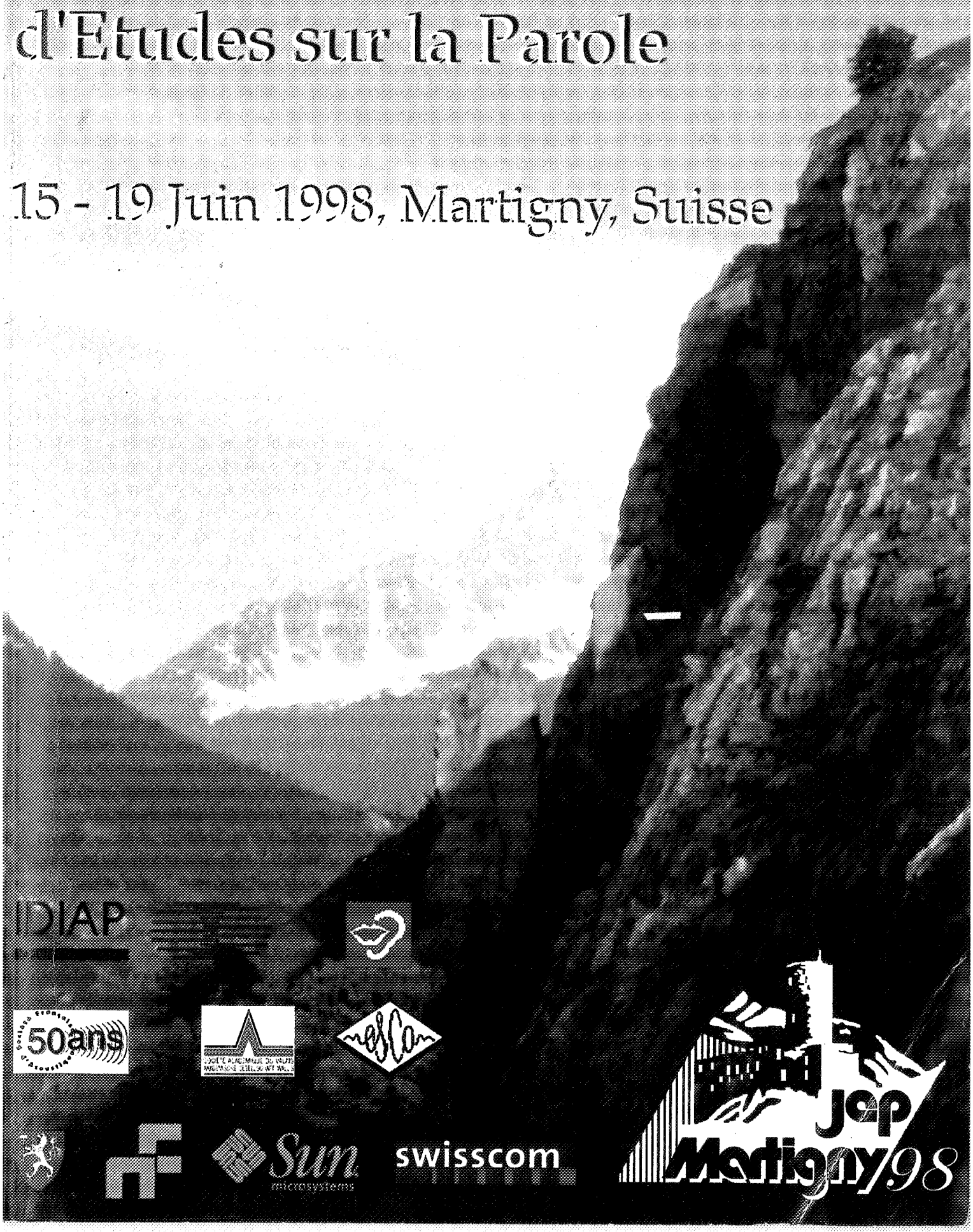
Actes des XXIIèmes Journées d'Etudes sur la Parole

15 - 19 Juin 1998, Martigny, Suisse



CERM

Centre d'Expositions et de Réunions, Martigny

XXIIèmes

JOURNEES D'ETUDE SUR LA PAROLE

(JEP'98)

du 15 au 19 juin 1998

Martigny - Valais - Suisse

organisées par

l'IDIAP - Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Suisse

pour

le GFCP - Groupe Francophone de la Communication Parlée

(Groupe spécialisé de la SFA)

avec le soutien de

la SFA - Société Française d'Acoustique

la SAV - Société Académique du Valais

l'ESCA - European Speech Communication Association

et le parrainage de

FNRS - Fonds National Suisse pour la Recherche Scientifique

SUN Microsystems

la Ville de Martigny

SWISSCOM

XXIIèmes
JOURNEES D'ETUDE SUR LA PAROLE

COMITÉ SCIENTIFIQUE

M. ADDA-DECKER, LIMSI, Paris (F)
R. ANDRÉ-OBRECHT, IRIT, Toulouse (F)
F. BIMBOT, IRISA, Rennes - ENST, Paris (F)
J.-F. BONASTRE, LIA, Avignon (F)
H. BOURLARD, IDIAP, Martigny (CH)
J.-L. COCHARD, SWISSCOM, Berne (CH)
P. DUPONT, EURISE, St-Etienne (F)
P. DELÉGLISE, LIUM, Le Mans (F)
J.-M. HOMBERT, DDL, Lyon (F)
Y. LAPRIE, LORIA, Nancy (F)
R.K. MOORE, DERA, Malvern (UK)
C. MONTACIÉ, LIP6, Paris (F)
P. PERRIER, ICP, Grenoble (F)
J. SCHOENTGEN, ULB, Bruxelles (B)
R. SOCK, IPS, Strasbourg (F)
B. TESTON, LPL, Aix-en-Provence (F)
J. VAISSIÈRE, IPP, Paris (F)

COMITÉ D'ORGANISATION

H. BOURLARD (Président), C. PILLET (Secrétaire),
O. BORNET, G. CALOZ, F. FORMAZ, D. GENOUD, S. KRSTULOVIC, J. MARIETHOZ.

MESSAGE DE BIENVENUE DU PRÉSIDENT DU COMITÉ D'ORGANISATION

Au nom du Comité d'Organisation des JEP'98, je suis heureux de vous souhaiter la bienvenue aux XXIIèmes Journées d'Etude sur la Parole et de vous accueillir à Martigny, site romain et carrefour renommé de tout temps des grandes voies de communication à travers les Alpes.

J'espère que cette conférence sera, elle aussi, une bonne voie de communication entre les nombreuses disciplines relatives au traitement de la parole représentées ici. Je saisis cette occasion pour remercier tous nos collègues qui ont soumis une contribution à cette conférence, ainsi que les membres du Comité Scientifique pour leur sérieux et leur dévouement lors de la sélection des communications.

Depuis ces dernières années, le domaine du traitement de la parole a été le théâtre d'une très grande activité, tant au niveau industriel que scientifique. Alors que le nombre de sociétés et d'applications dans ce domaine a fortement augmenté, la recherche scientifique se remet doucement en question et devient de plus en plus multi-disciplinaire, utilisant des résultats provenant de domaines aussi divers que les mathématiques et les statistiques, la linguistique et la théorie du langage, la neurobiologie et la psycho-acoustique. Depuis leur création, les JEP ont eu à coeur de refléter cette multi-disciplinarité, toujours aussi présente cette année. Evidemment, ceci conduit aussi à des travaux de recherche de plus en plus complexes et difficiles à appréhender, rendant parfois leur accès difficile à nos jeunes chercheurs. Faciliter cet accès a toujours été un des soucis majeurs des JEP. Malgré le caractère de plus en plus international et exigeant de la recherche en général, une conférence telle que les JEP a toujours son rôle à jouer en reflétant les grandes tendances internationales, tout en y intégrant certaines spécificités plus «locales».

Les JEP offrent également l'occasion de rencontrer de nombreux collègues dans un environnement agréable et convivial. Parmi les aspects moins techniques de ces JEP'98, j'espère aussi que chacun de vous trouvera l'occasion d'apprécier certains échantillons représentatifs de Martigny, capitale artistique du Valais, haut lieu gastronomique et pays de la Williamine, née de la bise et du foehn, et dont les habitants se fondent au rythme de ces vents qui les imprègnent de leur contradiction. Evidemment, vous êtes tous les bienvenus à notre soirée valaisanne du Mercredi 17 Juin, et nous vous recommendons vivement de vous éloigner pour quelques heures de vos occupations scientifiques afin de visiter la Fondation Gianadda qui accueille actuellement une exposition exceptionnelle consacrée à Gauguin.

Bienvenue donc en Valais, région touristique située au coeur des Alpes, aux stations de ski de renommée internationale, mais qui proposent également à ses hôtes la découverte de ses sentiers pittoresques, ses vins typiques (Fendant, Gamay, Arvine, Cornalin), ses villages imprégnés de tradition au coeur d'une vallée où le Rhône a tracé son parcours.

Evidemment, cette conférence n'aurait pas vu le jour sans le soutien du GFCP (Groupe Francophone de la Communication Parlée) de la SFA (Société Française d'Acoustique) laquelle fête cette année ses 50 ans, ainsi que de l'ESCA (European Speech Communication Association). Nous remercions également les institutions qui ont accepté d'apporter un soutien financier à cette conférence, à savoir : le Fonds National Suisse de la Recherche Scientifique, la Société Académique du Valais, Sun Microsystems, et SWISSCOM.

Finalement, je n'insisterai jamais assez pour remercier très sincèrement et très cordialement les membres du Comité d'Organisation pour le travail remarquable qu'ils ont accompli, souvent en l'absence de leur Président. Tout le mérite du succès de cette conférence leur revient !

Hervé Bourlard
Président du Comité d'Organisation

MESSAGE DE BIENVENUE DU PRÉSIDENT DU GFCP

Ce 15 juin 1998, les 22èmes Journées d'Etude sur la Parole vont s'ouvrir, pour 5 jours. C'est l'IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive) situé à Martigny, en Suisse, qui accueille cette manifestation.

La tenue régulière des JEP depuis maintenant plus de 25 ans sous l'égide du Groupe Francophone de la Communication Parlée (GFCP) de la SFA (Société Française d'Acoustique) montre bien, si tant est que ce soit nécessaire, que cette manifestation correspond à un besoin bien précis de la communauté scientifique francophone travaillant sur la communication parlée : celui de se retrouver à intervalles réguliers pour échanger, en langue française, les résultats de travaux scientifiques récents. En effet, une telle conférence est complémentaire des événements européens et internationaux qui permettent eux aussi de faire connaître nos travaux au reste de la communauté scientifique avec cependant la contrainte de devoir s'exprimer dans une langue différente de celle dans laquelle ces travaux ont été le plus souvent conçus et développés. Aussi le caractère francophone des JEP permet-il de conserver un espace de plus grande spontanéité et de plus grande convivialité.

Comme d'habitude, bon nombre de pays francophones sont représentés à côté de la France : la Suisse, bien sûr, mais aussi l'Algérie, la Belgique et le Canada. Il faut également noter la présence de communications émanant partiellement ou totalement de laboratoires non-francophones, qu'il s'agisse de résultats de collaborations entre laboratoires francophones et non-francophones, de communications co-signées par un chercheur francophone séjournant dans un laboratoire non-francophone ou encore d'articles dont les auteurs, non-francophones, ont néanmoins choisi le français comme langue d'expression pour présenter leurs travaux. A ces titres divers, vous trouverez donc dans ces actes des auteurs affiliés à des laboratoires en Allemagne, en Colombie, en Corée du Sud, aux Etats-Unis, en Italie, au Japon, aux Pays-Bas, en République Tchèque et en Suède. Cette diversité géographique est le signe que les JEP continuent à susciter l'intérêt de la communauté scientifique au-delà des limites strictes de la francophonie.

Parallèlement, les JEP constituent aussi une occasion pour de nombreux jeunes chercheurs de présenter leurs travaux à leurs aînés et de bénéficier de leur conseils, dans un contexte qui facilite les interactions. Pour beaucoup d'entre nous, c'est aux JEP que nous avons, pour la première fois, exposé nos recherches dans le cadre d'une conférence scientifique. Les JEP remplissent donc aussi une fonction quasiment "initiatique", et c'est certainement une raison supplémentaire de leur succès au fil des ans.

Depuis maintenant plusieurs années, les communications pour les JEP sont sélectionnées sur la base d'un article complet. Cela permet d'opérer une évaluation bien plus précise de la qualité des propositions de communications. Du même coup, le nombre de propositions d'articles est un peu moins élevé, mais leur niveau est globalement meilleur. En outre, cette année, certaines communications ont été retenues sous réserve de modifications impératives (environ 30 % des articles soumis) : il s'agissait en général d'articles de bonne qualité globale, mais pour lesquels un point précis nécessitait d'être explicité ou rédigé différemment. Pour ces textes, une seconde relecture a été mise en place, afin de s'assurer que les demandes des relecteurs avaient effectivement été prises en compte par les auteurs. Cette procédure a certainement été bénéfique aux articles soumis à révision tout en évitant d'avoir à effectuer une décision trop binaire lors de la première lecture. Au total, ce cru des JEP contient 89 communications.

Les actes débutent par une section spéciale en mémoire de Christian Benoît qui nous a brutalement quittés à la fin du mois d'avril dernier. Christian Benoît était un chercheur d'envergure internationale comme l'attestent le volume et la qualité de sa production scientifique et son intense participation à la vie de la communauté à l'échelle nationale et supra-nationale. Il tenait néanmoins à publier régulièrement dans les JEP, en français, ses travaux les plus récents. Le GFCP a donc décidé de reproduire l'intégrale des articles qu'il a publié dans cette conférence tout au long de sa carrière, en hommage au chercheur francophone de renommée internationale qu'il était.

La suite du volume contient les 89 communications retenues, dont une petite moitié sera présentée sous forme de sessions affichées. Le choix du mode de présentation n'ayant été effectué que sur des critères strictement thématiques, la distinction n'apparaît pas dans les actes. Par ailleurs, nous avons maintenu le principe de communications orales uniquement sous forme de sessions plénières, afin de permettre à tous les participants d'assister à tous les exposés. C'est, à nos yeux, une façon de favoriser l'inter-disciplinarité de cette manifestation.

Ces actes contiennent également les textes de 4 conférences invitées sur des thèmes aussi divers que la dynamique des systèmes phonologiques, l'état émotionnel du locuteur, les ondelettes en traitement de la parole et les systèmes de

dialogue oral. Nous avons par ailleurs réservé un créneau horaire particulier pour une table ronde sur les ressources linguistiques, organisée par ELRA (European Language Resources Association / Association Européenne pour les Ressources Linguistiques), et précédée d'une conférence d'introduction dont le texte figure également dans les actes. Ces créneaux spécifiques permettront aux participants d'approfondir certains sujets qui nous ont semblé particulièrement d'actualité.

Je terminerai en remerciant l'ensemble des membres du Comité d'Organisation qui ont patiemment mis en place toutes les pièces du puzzle et ont rendu possible la tenue de cette dernière édition des JEP avant l'an 2000. Compte tenu de la légendaire précision suisse, et ayant pu moi-même faire l'expérience de l'hospitalité des valaisans (et de ceux qui ont la chance de travailler avec eux), je suis convaincu que la qualité de l'organisation et la chaleur de l'ambiance de ces JEP resteront longtemps présents à l'esprit de ceux qui y auront participé.

Frédéric BIMBOT
Président du GFCP

TABLE DES MATIÈRES

Hommage à Christian Benoît

Locuteur I

Système d'élagage temps-fréquence pour l'identification du locuteur
(L. Besacier, J.F. Bonastre) 1
Informations dynamiques et méthodes statistiques du second ordre pour l'identification du locuteur
(C. Fredouille, J.F. Bonastre) 5
Vérification du locuteur dépendante du texte utilisant des classificateurs binaires
(D. Genoud, M. Moreira, E. Mayoraz) 9

Locuteur II

Reconnaissance acoustique et perceptuelle des ressemblances vocales familiales
(E. Perrin, L. Collet, C. Berger-Vachon) 13
Méthode d'estimation du seuil de décision en vérification du locuteur
(J.B. Pierrot, F. Bimbot) 17
Proposition d'une stratégie de fusion de données à trois niveaux pour la vérification d'identité
(P. Verlinde, D. Genoud, G. Gravier, G. Chollet) 21

Psycholinguistique

Les opérations pragmatiques d'une communication agrammatique
(F. Argod-Dutard) 25
Contribution des indices métriques à l'apprentissage d'un langage artificiel
(M.H. Banel, U.H. Frauenfelder, P. Perruchet) 29
Le rôle de la syllabe : segmentation lexicale ou classification ?
(N. Dumay, M.H. Banel, U.H. Frauenfelder, A. Content) 33
Rôle de la médiation phonologique dans l'apprentissage de la lecture
(G. Fauzar, M. Farid) 37
Segmentation en mots et compétitions lexicales
(F. Isel, N. Bacri) 41
Sensibilité à la contingence phonémique et à la fréquence des unités infra-syllabiques dans les jugements phonotactiques (R. Peereman, E. Ferreri, P. Perruchet) 45
Facilitation initiale en amorçage phonologique : effet syllabique ou phonémique ?
(E. Spinelli, J. Segui) 49

Conférence invitée

Dynamique des systèmes phonologiques - Aspects formels et expérimentaux
(D. Demolin) 53

Phonétique I

Identification dialectale des parlers arabes et détermination expérimentale d'indices discriminants
(M. Barkat) 71
Classement automatique de phonèmes dans un cadre multilingue
(C. Corredor-Ardoy, P. Boula de Mareüil, M. Adda-Decker, L. Lamel, J.L. Gauvain) 75
Production de phones « nouveaux » et « similaires » en L2
(H.Z. Kim) 79
La partition de l'espace vocalique en berbère rifain
(N. Louali, G. Puech) 83
La phonétisation des emprunts et de leur paradigme flexionnel
(F. Sannier, R. Belrhali, V. Aubergé) 87
Organisation temporelle des voyelles dans les structures CVCVCV, CVCCVCV et CVCCV de l'arabe standard
(A. Amrouche, B. Boudraa, J.M. Rouvaen) 91

Prosodie

Analyse multi-paramétrique des configurations tonales des proéminences en français
(C. Astesano, R. Espesser, E. Flachaire, P. Nicolas) 95
Evaluation de modèles d'étiquetage automatique de l'intonation
(E. Campione, E. Flachaire, D. Hirst, J. Véronis) 99

Les critères prosodiques des trois principaux types de phrases (testés sur le tchèque synthétique)
(M. Dohalská, J. Mejvaldová) 103
Expériences intonatives sur l'intonation
(D. Duvivier) 107
Pondération d'indices prosodiques dans un système automatique de prédiction syntaxico-rythmique d'un énoncé
(P. Langlais, J.L. Cochard) 111
Le même est-il un autre ? Indices prosodiques de désambiguïsation de certaines répétitions à l'oral
(C. Leleu) 115
Association prosodie-syntaxe : validation par synthèse
(P. Martin) 119
Mesure perceptive de la fonction linguistique de la prosodie pure à l'aide d'énoncés réitérés
(A. Rilliard, V. Aubergé) 123
L'intonation de la variété francoprovençale parlée à Sarre (Aoste)
(S. Roullet, A. Romano) 127
Caractérisation du débit de parole en français
(B. Zellner) 131

Synthèse

Optimisation d'un modèle prosodique pour la synthèse par règles à partir du texte en français
(A. Di Cristo, P. Di Cristo, J. Véronis) 135
Le système de synthèse FIPSVox : syntaxe, phonétisation et prosodie
(A. Gaudinat, J.P. Goldman) 139
Une méthode de synthèse par copie pour piloter le synthétiseur de Klatt
(Y. Laprie, A. Bonneau) 143
Un générateur de prosodie « tout automatique »
(F. Malfrère, T. Dutoit, P. Mertens) 147
Utilisation d'une structure arborescente pour une hiérarchisation fine des règles de transcription
graphème-phonème (M. Morel, A. Lacheret-Dujour) 151
Contour intonatif et syntaxe en arabe : résultats préliminaires
(Z. Najim) 155

Parole et émotion

Perception de l'émotion : les indices prosodiques et la durée minimale
(S.J. Chung) 159
Pertinence perceptive des configurations intonatives en parole émotionnelle
(S.J.L. Mozziconacci, D.J. Hermes) 163
La variabilité du rythme cardiaque et la différentiation prosodique des émotions
(B. Zei, M. Archinard) 167

Ressources

Spécification, réalisation et validation d'un corpus oral pour la reconnaissance de la parole dans une voiture
(C. Gassert, J.F. Mari) 171
EMACOP : Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole
(D. Vaufreydaz, M. Akbar, J. Caelen, J.F. Sérignat) 175

Langue

Indentification des langues par discrimination automatique des systèmes vocaliques
(F. Pellegrino, A. Galinier) 179

Table ronde : conférence d'introduction

L'Association Européenne pour les Ressources Linguistiques (ELRA)
(K. Choukri, V. Mapelli) 183

Conférence invitée

Analyse en ondelettes et reconnaissance de la parole
(S.H. Maes, T.J. Watson) 187

Parole audiovisuelle

Perception visuelle du glide d'arrondissement en français
(M.A. Cathiard) 201

Interactions audiovisuelles dans la perception de la parole en français
(C. Colin, M. Radeau, P. Deltenre) 205
Codage audiovisuel de la parole : une étude préliminaire de la réduction du débit d'un vocodeur LPC par utilisation de paramètres vidéo (E. Foucher, G. Feng, L. Girin) 209
Un modèle géométrique de lèvres 3D adaptable au locuteur
(L. Revéret, L. Le Chevalier) 213
Apports d'un apprentissage discriminant du visuel à la reconnaissance audiovisuelle de la parole
(A. Rogozan, P. Deléglise) 217
Reconnaissance audiovisuelle de voyelles bruitées : estimation de la fiabilité du capteur audio et contrôle du processus de fusion (P. Teissier, J.L. Schwartz, A. Guérin-Dugué) 221

Phonétique II

L'étude des voyelles [*a*] et [a] en français et en persan
(S. Assadi) 225
Variabilité intra-locuteur dans la production du phonème /R/ en français
(M. Chafcouloff, D. Autesserre) 229
Analyse phonétique des séquences /tch/ et /dz/ en français
(C. Corneau) 233
Etude acoustique préliminaire des 16 voyelles orales du Paez de Talaga, langue amérindienne
(E. Marsico, T. Rojas Curieux) 237
Les systèmes consonantiques. Des tendances universelles à l'ontogenèse
(N. Vallée, L.J. Boë, M. Stefanuto) 241
[i a u] ? A la portée d'un conduit vocal de Néandertal
(L.J. Boë, S. Maeda, C. Abry, J.L. Heim) 245

Conférence invitée

L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole
(K. R. Scherer, T. Johnstone, J. Sangsue) 249

Perception

Effet du renforcement des modulations temporelles en sous-bandes sur la reconnaissance de la parole : Résultats préliminaires (F. Apoux, F. Berthommier, N. Bacri, C. Lorenzi) 259
Perception des gestes de parole
(R. Carré, P.L. Divenyi) 263
Est-il absurde d'entendre /b/ dans « absurde »
(C. Chéreau, P. Halle, J. Segui) 267
Nature des représentations prosodiques en mémoire auditive
(J. Clément, C. Gérard) 271
Groupes consonantiques illégaux et segmentation de la parole
(O. Crouzet, N. Bacri) 275
Unité minimale de perception de la parole chez le nouveau-né : capacité de discrimination syllabique ou moraïque ? (H. Yamasaki) 279

Production

Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM (P. Badin, L. Pouchoy, G. Bailly, M. Raybaudi, C. Segebarth, J.F. Lebas, M. Tiede, E. Vatikiotis-Bateson, Y. Tohkura) 283
Contribution de la mâchoire à la géométrie de la langue dans les modèles articulatoires statistiques
(G. Bailly, P. Badin, A. Vilain) 287
Articulation consonantique en début de constituants prosodiques de différents niveaux
(C. Fougeron) 291
Inversion acoustique articulatoire par une méthode variationnelle
(Y. Laprie, B. Mathieu) 295
Relation entre pression sous glottique et intensité : étude des voyelles du français
(V. Lecuit, D. Demolin) 299
Frontière syllabique, accent focal contrastif et coarticulation linguale dans le groupe consonantique /kskl/ en position médiane de mot en français (Y. Meynadier, A. Marchal, M. Pitermann) 303
La relation entre les niveaux articulatoire et acoustique peut être décrite par des équations différentielles couplées
(J. Schoentgen, S. Ciocea) 307
A propos des degrés de liberté dans la coarticulation d'un locuteur français filmé 50 fois par seconde
(A. Vilain, C. Abry, P. Badin) 311

Analyse et codage

Codage de canal à source dépendante : application au codeur « CELP »
(B. Boudraa, M. Bouzid, M. Boudraa, B. Guérin) 315
Filtrages logiques adaptés au pitch
(M. Boudraa, H. Sayoud, B. Boudraa, B. Guérin) 319
Critère statistique pour l'automate de détection bruit/parole
(L. Karray, J. Monné) 323
Utilisation des modèles de Markov cachés pour le débruitage
(D. Matrouf, J.L. Gauvain) 327
Classification parole/musique de signaux sonores paramétrisés par les composantes principales du spectre
(M. Seck, F. Bimbot, B. Delyon) 331
Utilisation de contraintes articulatoires pour l'extraction automatique des formants
(A. Soquet, V. Lecuit) 335
Décomposition spectro-temporelle de syllabes et de mots superposés : résultats préliminaires
(L. Varin, F. Berthommier) 339

Modèles de langage

Un modèle de langage mixte basé sur la similarité des mots dans un système de reconnaissance de parole
(C. Beaujard, M. Jardino) 343
Combinaison de modèles de langages pour l'identification de thèmes
(B. Bigi, R. De Mori, M. El-Bèze, T. Spriet) 347
Apport de l'utilisation de classes de mots pour la définition d'un nouveau lissage de type absolute discounting
(G. Damnati, J. Simonin) 351
Modélisation statistique du langage avec un modèle bi-multigramme
(S. Deligne, Y. Sagisaka) 355
L'analyse syntaxique robuste de dialogues transcrits : de la parole aux actes en passant par l'écrit
(S. Delisle, N. Boufaden, B. Moulin) 359
Evaluation de modèles de langage à base de trigrammes de classes et de mots, avec le Jeu de Shannon
(M. Jardino) 363

Reconnaissance I

Eléments pour la mise au point de système de reconnaissance grand vocabulaire en français
(M. Adda-Decker, G. Adda, J.L. Gauvain, L. Lamel) 367
Etude de la recombinaison de plusieurs classifieurs appliquée à deux tâches de reconnaissance de la parole
(C. Cerisara, M. Afify, J.P. Haton) 371
Reconnaissance multi-bandes de la parole bruitée par couplage entre niveaux primitif et d'identification
(H. Glotin, E. Tessier, H. Bourlard, F. Berthommier) 375
Reconnaissance de noms prononcés et épelés au téléphone par modèles de Markov cachés
(F. Neubert, G. Gravier, F. Yvon, G. Chollet) 379
Indexation vocale à vocabulaire illimité à base de décodage phonétique
(S. Peillon, A. Ferrieux) 383

Conférence invitée

Spoken language dialogue : architectures and algorithms
(R. Pieraccini, E. Levin, W. Eckert) 387

Reconnaissance II

Reconnaissance des phonèmes par réseaux de neurones et normalisation temporelle : application aux consonnes pharyngales et glottale arabes (A. Amrouche, M. Debyeche, A. Adoul, K. Amrouch, J.M. Rouvaen) 397
Modèles hybrides et reconnaissance de la parole continue indépendante du locuteur en français
(O. Deroo, C. Ris, F. Malfrère, T. Dutoit) 401
Reconstruction de données manquantes pour la reconnaissance robuste de la parole dans le cadre des systèmes hybrides HMM/ANN (S. Dupont) 405
Rehaussement par soustraction spectrale et compensation des paramètres manquants pour la reconnaissance robuste du locuteur et de la parole (M. El-Maliki, P. Renevey, A. Drygajlo) 409
La détection de nouveauté basée sur le temps de stabilisation d'un réseau de neurones : application possible en reconnaissance de parole ? (J. Rouat, T.V. Ho) 413
Identification de traits phonétiques arabes par des systèmes connexionnistes modulaires
(S.A. Selouani, J. Caelen) 417

INDEX DES AUTEURS

Abry C., 245, 311
Adda G., 367
Adda-Decker M., 75, 367
Adoul A., 397
Afify M., 371
Akbar M., 175
Amrouch K., 397
Amrouche A., 91, 397
Apoux F., 259
Archinard M., 167
Argod-Dutard F., 25
Assadi S., 225
Astesano C., 95
Aubergé V., 87, 123
Autesserre D., 229
Bacri N., 41, 259, 275
Badin P., 283, 287, 311
Bailly G., 283, 287
Banel M.H., 29, 33
Barkat M., 71
Beaujard C., 343
Belrhali R., 87
Berger-Vachon C., 13
Berthommier F., 259, 339, 375
Besacier L., 1
Bigi B., 347
Bimbot F., 17, 331
Boë L.J., 241, 245
Bonastre J.F., 1, 5
Bonneau A., 143
Boudraa B., 91, 315, 319
Boudraa M., 315, 319
Boufaden N., 359
Boula de Mareüil P., 75
Bourlard H., 375
Bouzid M., 315
Caelen J., 175, 417
Campione E., 99
Carré R., 263
Cathiard M.A., 201
Cerisara C., 371
Chafcouloff M., 229
Chéreau C., 267
Chollet G., 21, 379
Choukri K., 183
Chung S.J., 159
Ciocea S., 307
Clément J., 271
Cochard J.L., 111
Colin C., 205
Collet L., 13
Content A., 33
Corneau C., 233
Corredor-Ardoy C., 75
Crouzet O., 275
Damnati G., 351
De Mori R., 347
Debyeche M., 397
Deléglise P., 217
Deligne S., 355
Delisle S., 359
Deltenre P., 205
Delyon B., 331
Demolin D., 53, 299
Deroo O., 401
Di Cristo A., 135
Di Cristo P., 135
Divenyi P.L., 263
Dohalská M., 103
Drygajlo A., 409
Dumay N., 33
Dupont S., 405
Dutoit T., 147, 401
Duvivier D., 107
Eckert W., 387
El-Bèze M., 347
El-Maliki M., 409
Espesser R., 95
Farid M., 37
Fauzar G., 37
Feng G., 209
Ferreri E., 45
Ferrieux A., 383
Flachaire E., 95, 99
Foucher E., 209
Fougeron C., 291
Frauenfelder U.H., 29, 33
Fredouille C., 5
Galinier A., 179
Gassert C., 171
Gaudinat A., 139
Gauvain J.L., 75, 327, 367
Genoud D., 9, 21
Gérard C., 271
Girin L., 209
Glotin H., 375
Goldman J.P., 139
Gravier G., 21, 379
Guérin B., 315, 319
Guérin-Dugué A., 221
Halle P., 267
Haton J.P., 371
Heim J.L., 245
Hermes D.J., 163
Hirst D., 99
Ho T.V., 413
Isel F., 41
Jardino M., 343, 363
Johnstone T., 249
Karray L., 323
Kim H.Z., 79
Lacheret-Dujour A., 151
Lamel L., 75, 367
Langlais P., 111
Laprie Y., 143, 295
Le Chevalier L., 213
Lebas J.F., 283
Lecuit V., 299, 335
Leleu C., 115
Levin E., 387
Lorenzi C., 259
Louali N., 83
Maeda S., 245
Maes S.H., 187
Malfrère F., 147, 401
Mapelli V., 183
Marchal A., 303
Mari J.F., 171
Marsico E., 237
Martin P., 119
Mathieu B., 295
Matrouf D., 327
Mayoraz E., 9
Mejvaldová J., 103
Mertens P., 147
Meynadier Y., 303
Monné J., 323
Moreira M., 9
Morel M., 151
Moulin B., 359
Mozziconacci S.J.L., 163
Najim Z., 155
Neubert F., 379
Nicolas P., 95
Peereman R., 45
Peillon S., 383
Pellegrino F., 179
Perrin E., 13
Perruchet P., 29, 45
Pieraccini R., 387
Pierrot J.B., 17
Pitermann M., 303
Pouchoy L., 283
Puech G., 83
Radeau M., 205
Raybaudi M., 283

Renevey P., 409
Revéret L., 213
Rilliard A., 123
Ris C., 401
Rogozan A., 217
Rojas Curieux T., 237
Romano A., 127
Rouat J., 413
Roullet S., 127
Rouvaen J.M., 91, 397
Sagisaka Y., 355
Sangsue J., 249
Sannier F., 87
Sayoud H., 319
Schoentgen J., 307
Schwartz J.L., 221
Seck M., 331
Segebarth C., 283
Segui J., 49, 267
Selouani S.A., 417
Sérignat J.F., 175
Simonin J., 351
Soquet A., 335
Spinelli E., 49
Spriet T., 347
Stefanuto M., 241
Teissier P., 221
Tessier E., 375
Tiede M., 283
Tohkura Y., 283
Vallée N., 241
Varin L., 339
Vatikiotis-Bateson E., 283
Vaufreydaz D., 175
Verlinde P., 21
Véronis J., 99, 135
Vilain A., 287, 311
Watson T.J., 187
Yamasaki H., 279
Yvon F., 379
Zei B., 167
Zellner B., 131

ERRATA

Dans l'Index des Auteurs

Manque: Scherer K., 249

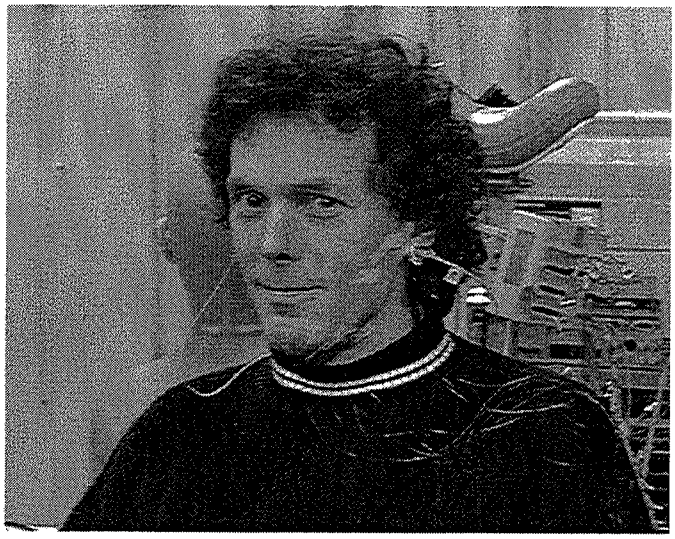
Pour l'article "Identification dialectale des parlers arabes et détermination expérimentale d'indices discriminants (*M. Barkat*)" p. 72

dans la partie 4.1.2. "Identification par pays et première approche d'analyse des erreurs" il convient de lire :

[...] "Toutefois, il convient de remarquer que les sujets ***moyen-orientaux*** présentent un taux d'identification par pays homogène (i.e. bonne localisation des parlers maghrébins et moyen-orientaux) par rapport aux sujets ***maghrébins***, ces derniers distinguent, en effet, assez mal les parlers ***moyen-orientaux*** entre eux."

Hommage à Christian Benoît

Hommage à Christian Benoît



Christian Benoît était un inconditionnel des JEP. Depuis 1984, année du début de sa thèse de Docteur Ingénieur, il n'a pas manqué une seule édition de ce congrès d'abord annuel, puis bi-annuel à partir de 1988. Il en était fier, et ne manquait pas de le rappeler à ceux qui, ne jurant que par les grandes manifestations internationales, avaient tendance à bouder ces rencontres francophones. Car Christian, ardent défenseur de la francophonie, tout en étant omniprésent sur la scène internationale, défendait avec vigueur, en ces Journées, l'idée d'un lieu privilégié, où des échanges scientifiques de qualité pouvaient se faire en langue française et autour de la langue française. Profitant de cette possibilité unique de s'exprimer et d'écrire dans notre langue qu'il maniait avec brio et nuance, il a fait preuve d'une belle productivité, publiant 16 articles dans les actes de ces Journées, avec une régularité qui n'a été mise en défaut qu'en 1988. Du "*Quelques exigences venues d'on ne sait haut pour la lecture des plans factoriels en Phonétique.*" en 1984, au titre de son article de cette année, en passant par "*L'organisation temporelle CVCV chez quatre locuteurs français ou faut-il chercher l'invariant dans les VC*" et par "*De la production à l'extraction, l'état d'un chantier.*" Christian Benoît a montré tout à la fois l'étendue de ses compétences de chercheur, et son art à manier l'humour et la provocation tout en exposant ses résultats avec sérieux et précision.

Entre ses premiers travaux sur l'analyse temporelle de la parole et ses recherches plus récentes sur l'intégration audition-vision dans la perception de la parole, Christian Benoît a réalisé un éditeur du signal de parole, a approfondi certains aspects de l'analyse de données, a travaillé sur l'évaluation de synthétiseurs de parole et a contribué au développement de la synthèse audiovisuelle de la parole, couvrant ainsi un vaste domaine dans le champ des recherches sur la Communication Parlée. Il était devenu logiquement un élément moteur de notre communauté scientifique, et il était incontestablement une des grandes figures internationales dans le domaine de la parole audiovisuelle, dont il affectionnait particulièrement le caractère pluridisciplinaire associant ingénieurs, linguistes et psychologues. Il est l'auteur ou le co-auteur de plus de 70 publications. Il était membre de l'*Advisory Board* et de l'*Editorial Board* du journal *Speech Communication*. Il avait organisé ou contribué à organiser de nombreuses rencontres et symposiums sur la parole audiovisuelle, et il était sur le point de créer, au sein d'ESCA, d'une société internationale consacrée

à ce domaine, AVISA. Il était impliqué dans de très nombreuses collaborations internationales, associant des laboratoires du monde entier, en Europe, en Amérique du Nord, en Asie.

Christian Benoît joua aussi un rôle important au sein de nos instances, nationales et internationales. Membre du Bureau du GCP (le futur GFCP) dès 1987, il en sera un Secrétaire actif, jusqu'en 1992. À ce titre, il participa au comité scientifique des Journées d'Études sur la Parole, et créa le fameux *Tchatch Comm'*, bulletin de liaison du groupe, qu'il animera avec dynamisme. Ce bulletin fut stoppé en 1991, et n'a aujourd'hui pas de remplaçant. Il s'est investi ensuite très activement au sein de l'ESCA, dont il intégra le Bureau en 1993. Il en a été le Secrétaire-Trésorier de 1995 à 1997, et il avait été reconduit dans ses fonctions de Secrétaire lors de la dernière assemblée générale à Rhodes en septembre dernier.

Christian Benoît venait de soutenir son Habilitation à Diriger des Recherches intitulée "*intelligibilité de la parole audiovisuelle pour l'homme et la machine*". Soucieux de faire le lien entre le monde scientifique et la société, il avait participé à de nombreuses manifestations à destination du grand public, et, très récemment encore, il était intervenu dans le cadre de l'affaire PRIETO, pour signaler les dangers de l'utilisation des techniques de reconnaissance du locuteur à des fins juridiques. Il avait dans le même temps aidé deux jeunes chercheurs de l'ICP à créer leur propre entreprise à Grenoble, GANYMEDIA, qui développe des produits comme le casque caméra-micro pour le suivi de la parole audiovisuelle et des logiciels de contrôle pour l'animation de têtes parlantes. Il venait aussi de lancer l'idée d'un "téléphone complété", à l'usage des sourds, en vue de transmettre à bas débit, les gestes manuels et labiaux nécessaires à la communication de cette communauté.

Christian Benoît ne sera pas parmi nous cette année. Il est décédé, brutalement, en pleine force de l'âge et au sommet de ses capacités, le 26 avril dernier, nous laissant hébétés et désemparés. Il avait 41 ans. Nous serons nombreux à souffrir pendant ses journées de l'absence de sa silhouette filiforme toujours en mouvement, de ses interventions toujours remarquées, et de son entrain communicatif. Bien des hommages scientifiques suivront sans doute dans les mois qui viennent, en particulier à l'occasion des manifestations qu'il contribuait à organiser au moment de sa disparition. Nous voulons le remercier ici au nom de la communauté francophone de la Communication Parlée, pour tout ce qu'il nous a apporté au plan scientifique et au plan humain, au cours de sa vie si courte et pourtant si productive. C'est pourquoi, nous republions, en introduction à ces actes, l'intégralité de ses contributions aux Journées d'Étude sur la Parole, y compris l'article qu'il avait soumis cette année. Le Comité Scientifique lui avait demandé d'en modifier le titre, pour lui donner un caractère plus scientifique, tout en conservant le titre d'origine en sous-titre. Il avait manifesté son mécontentement, et nous étions en cours de négociation avec lui. Comme un dernier clin d'œil à notre ami Christian, nous avons choisi de publier la version du texte qu'il avait initialement proposée.

Nota Bene : La photo ci-dessus représente Christian Benoît au cours de la préparation d'une expérience de recueil de données cinématiques et EMG sur le visage, au cours de la production de la parole. Elle a été prise aux laboratoires ATR de Kyoto en janvier 1997. Nous remercions son auteur Takaaki Kuratate (de la société Toshiba et d'ATR) et Eric Vatikiotis-Bateson d'ATR de l'avoir aimablement mise à notre disposition.

QUELQUES EXIGENCES VENUES D'ON NE SAIT HAUT POUR LA LECTURE DES PLANS FACTORIELS EN PHONETIQUE

C. ABRY, C. BENOIT : Institut de Phonétique de Grenoble
Institut de la Communication Parlée, Grenoble, L.A. C.N.R.S. 368.

Utilisant conjointement, depuis quelques années, plusieurs variantes principales de l'analyse factorielle des données, pour nos besoins phonéticiens, nous avons pensé qu'il était temps de confronter le meilleur de leurs possibilités à quelques exigences.

Comme tout praticien, nous nous attendons en fait, devant un plan factoriel, à retrouver essentiellement deux choses : 1. Une correspondance avec notre intuition globale des données qui tient à une expérience, à des hypothèses. Et de ce côté-ci, la connaissance du signal, de sa production, comme celle de la phonologie y est pour quelque chose ... 2. Nos chères variables (articulatoires, acoustiques, etc.), en utilisant les possibilités d'interprétation des facteurs que nous offrent certaines de ces analyses.

Laissant de côté, pour aujourd'hui, l'insatisfaction que nous ressentons encore vis-à-vis des possibilités de retour aux variables, nous ne nous attacherons qu'à spécifier nos exigences de type 1.

EXIGENCES

Nos hypothèses sont de manière privilégiée discriminantes. Ceci est vrai en phonologie, où règne le principe de distinctivité entre classes phonémiques. Encore vrai en phonétique où l'on tiendra compte des classes allophoniques. Nos hypothèses discriminantes sont simples : elles vont des exigences de la phonologie à celles de la phonétique, en rappelant quelques principes de manifestation de la structure linguistique.

Etant donnée une opposition à n termes : 1. L'hypothèse phonologique veut que ces termes restent distincts. 2. Elle voudrait que ceux-ci restent distincts dans tous les contextes. 3. L'hypothèse phonétique sera que ces termes se distinguent mieux dans certains contextes que dans d'autres. Il existe en effet des contextes (segmentaux, prosodiques, ..., individuels et sociaux) plus facilitants que d'autres pour opposer les termes (contrastes maximaux et minimaux, ABRY & BOË, 1981-1982). A la limite, on conçoit qu'il y ait des contextes où ces termes se confondent (neutralisation). 4. Autre hypothèse phonétique, elle concerne les allophones. Il n'y a aucune raison pour penser qu'une variation contextuelle identique puisse avoir le même effet sur tous les termes de l'opposition. Certains d'entre eux sont en effet retrouvent dans le même ordre pour D.L. et pour P.C. Soit ANACOR > ANAFAC > ANADIS pour les D phonologiques); ANADIS > ANACOR > ANAFAC (pour les D allophoniques).

Dans un classement par rangs, ANACOR obtient 5 points, ANADIS 4 et ANAFAC 3. Cet ordre confirme ANACOR comme le plus apte à représenter les contraintes sur les manifestations de la structure phonique.

QU'EST-CE A DIRE ?

On connait la puissance de cette analyse factorielle des correspondances par la publicité qui lui a été faite, notamment dans les sciences humaines (BENZECRI, 1982). On savait déjà qu'elle est sans doute la plus "structuraliste" de toutes les analyses - ce que met en évidence la notion de profil comparé par la distance du CHI-2 à tous les autres profils, profils-éléments ou profils-variables. On s'attendait donc au respect de la structure phonologique. Mais on ne savait pas qu'elle pouvait aussi bien rendre compte de la structure phonétique.

En ce sens, l'analyse discriminante va peut être un peu vite en besogne en appliquant le principe de distinctivité sans tenir compte des relations structurales de l'ensemble, en oubliant qu'il s'agit d'un système. Et pour une approche relationnelle, on ne peut se limiter à une normalisation des paramètres (ANAFAC). Chercher le plan qui maximise les distances interclasses (en minimisant les dispersions intraclasses) ? Oui, mais en utilisant une distance structurale. C'est ce que réussit BENZECRI avec le CHI-2. C'est aussi ce qu'on pourrait méditer pour toute prédiction des espaces phonétiques à partir du principe de distinctivité maximale dans la voie ouverte par LILJENCRANTS et LINDBLOM (1972).

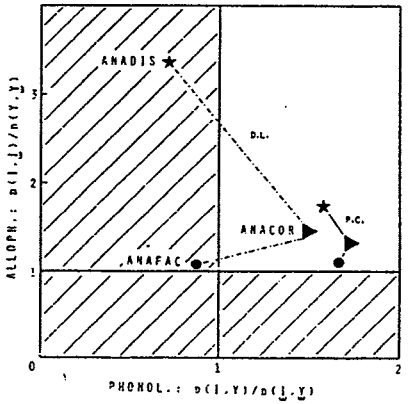


Figure 1

ABRY C. & al. (1980),
Labialité et phonétique. - Grenoble.

ABRY C. & BOË L.J. (1981-1982),
Sur les notions d'opposition et de contrastes ... - B.I.P.Grenoble, 10/11, 1-12.

BENOIT C. (1983),
Vers une herméneutique de l'analyse de données ... - B.I.P.Grenoble, 12, 155-183.

BENZECRI J.P. (1982),
Histoire et préhistoire de l'analyse des données. - Paris

LILJENCRANTS J. & LINDBLOM B. (1972),
Numerical simulation of vowel quality systems ... - Language 48, 839-862.

retrouvent dans le même ordre pour D.L. et pour P.C. Soit ANACOR > ANAFAC > ANADIS (pour les D phonologiques); ANADIS > ANACOR > ANAFAC (pour les D allophoniques).

Dans un classement par rangs, ANACOR obtient 5 points, ANADIS 4 et ANAFAC 3. Cet ordre confirme ANACOR comme le plus apte à représenter les contraintes sur les manifestations de la structure phonique.

QU'EST-CE A DIRE ?

On connait la puissance de cette analyse factorielle des correspondances par la publicité qui lui a été faite, notamment dans les sciences humaines (BENZECRI, 1982). On savait déjà qu'elle est sans doute la plus "structuraliste" de toutes les analyses - ce que met en évidence la notion de profil comparé par la distance du CHI-2 à tous les autres profils, profils-éléments ou profils-variables. On s'attendait donc au respect de la structure phonologique. Mais on ne savait pas qu'elle pouvait aussi bien rendre compte de la structure phonétique.

En ce sens, l'analyse discriminante va peut être un peu vite en besogne en appliquant le principe de distinctivité sans tenir compte des relations structurales de l'ensemble, en oubliant qu'il s'agit d'un système. Et pour une approche relationnelle, on ne peut se limiter à une normalisation des paramètres (ANAFAC). Chercher le plan qui maximise les distances interclasses (en minimisant les dispersions intraclasses) ? Oui, mais en utilisant une distance structurale. C'est ce que réussit BENZECRI avec le CHI-2. C'est aussi ce qu'on pourrait méditer pour toute prédiction des espaces phonétiques à partir du principe de distinctivité maximale dans la voie ouverte par LILJENCRANTS et LINDBLOM (1972).

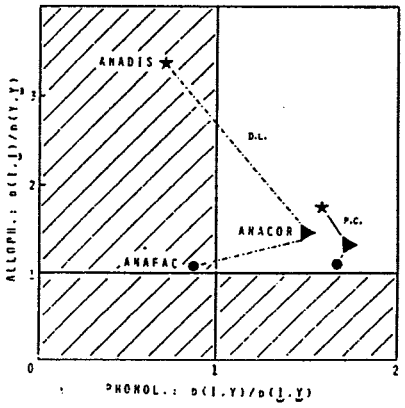


Figure 1

ABRY C. & al. (1980),
Labialité et phonétique. - Grenoble.

ABRY C. & BOË L.J. (1981-1982),
Sur les notions d'opposition et de contrastes ... - B.I.P.Grenoble, 10/11, 1-12.

BENOIT C. (1983),
Vers une herméneutique de l'analyse de données ... - B.I.P.Grenoble, 12, 155-183.

BENZECRI J.P. (1982),
Histoire et préhistoire de l'analyse des données. - Paris

LILJENCRANTS J. & LINDBLOM B. (1972),
Numerical simulation of vowel quality systems ... - Language 48, 839-862.

EDISIG : ENCORE UN EDITEUR DE SIGNAL ?!!

Ch. BENOIT : Institut de Phonétique de Grenoble
Institut de la Communication Parlée LA CNRS 368.

I - ATTAQUE DE LA BANQUE

Lorsqu'un Institut de Phonétique décide, dans le cadre de ses études* sur les caractéristiques individuelles et, plus largement, de ses recherches en acoustique, de constituer et d'analyser une banque de données**, il a besoin de pouvoir segmenter le corpus original, préalablement enregistré numériquement sur des bandes magnétiques, puis de l'étiqueter avant de s'attaquer enfin à l'extraction des paramètres qu'il souhaite étudier.

L'Institut de Phonétique de Grenoble et le LCP, déjà riches d'un héritage logiciel en visualisation et traitement du signal, avait cependant besoin d'un outil, sinon plus complet, au moins plus homogène et apte à concentrer en un même programme un ensemble de fonctions nécessaires au dépouillement d'une banque de données informatiques.

EDISIG tente de remplir ce rôle.

II - LIMITES DES EDITEURS EXISTANTS

A notre connaissance, aucun des différents programmes d'édition de signal déja réalisés et utilisés ne réunissent l'ensemble des caractéristiques requises :

- Souplesse d'utilisation et d'adaptation (langage type FORTRAN).
- Emploi d'un matériel "bon marché" et couramment utilisé (PDP 11 et carte de visualisation).
- Compatibilité avec des structures plus ou moins "standard" (type ILS).
- Utilisation de résultats acquis auparavant (lecture de fichiers calculés par ailleurs).
- Séparation totale du dialogue et du graphique (utilisation maximale de deux écrans).
- Utilisation dynamique des écrans par effacement partiel.
- Etc.

III - AVANTAGES D'EDISIG

Le créateur d'EDISIG, bien que n'étant pas informaticien, s'est fixé ces contraintes pour sa réalisation.

Le but du programme est de proposer le maximum de fonctions utiles à l'analyse d'une banque de données, depuis l'écoute du signal original jusqu'à la constitution de tableaux contenant les valeurs des différents paramètres étudiés, prêts à être soumis aux différentes méthodes d'analyse de données.

EDISIG, actuellement en cours de rédaction, permet déjà de :

- écouter le signal entre n'importe quelles bornes choisies,
- visualiser le signal(idem),
- segmenter les fichiers d'origine (format (I.L.S.) en petits fichiers de travail (compatibles I.L.S.),
- étiqueter ces différents fichiers,
- stocker les informations relatives au signal (références, étiquettes, etc.) dans des fichiers associés au fichier de signal,
- visualiser les courbes d'intensité et de FO du signal en synchronie avec celui-ci,
- calculer et visualiser le spectre instantané autour de l'échantillon de signal désiré,
- détecter manuellement les périodes ou les fréquences instantanées du signal,
- etc.

IV - FICHE TECHNIQUE

- Rédigé en FORTRAN, il nécessite :
  . Matériel PDP 11 + VT100 + carte ARINFO (512 x 512) + écran graphique + SCH (pour l'écoute exclusivement = CN/A).
  . Logiciel RT 11 + bibliothèque SYSLIB.
- Il construit à la segmentation les fichiers WD????.(signal) et leurs fichiers associés ET????. (description).
- Il utilise au besoin les fichiers annexes ID????.(intensité calculée extérieurement), FD????.(courbe F(t) idem.), SD????. (sonagramme numérique + détection de formants.
- Programmes complémentaires : calcul de l'intensité, de FO, du spectre et détection automatique des formants (FENG G., 1983).

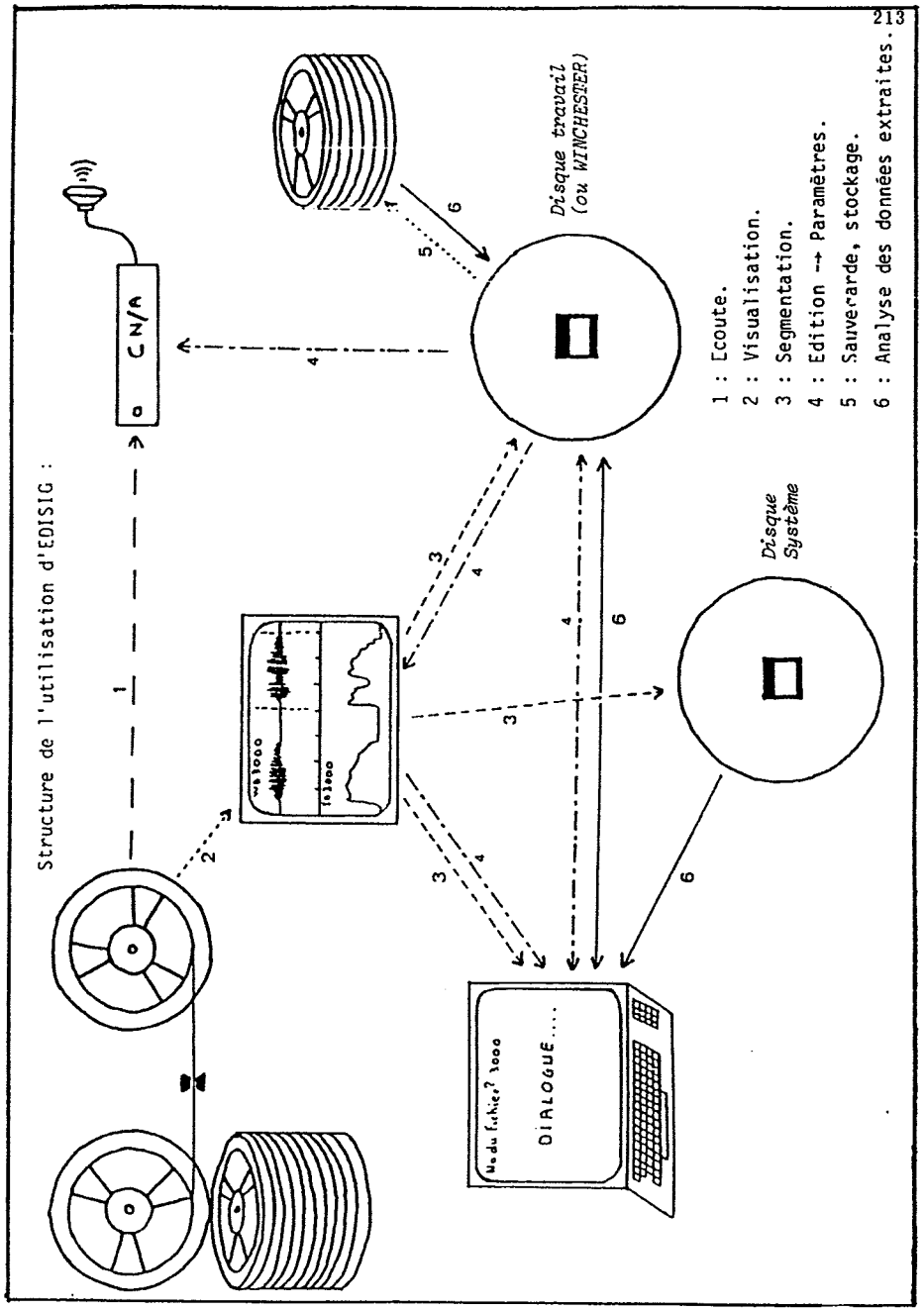
.....

FENG G. (1983)
Détection et mesure automatique de la fréquence fondamentale et des formants du signal de parole. - Rapport de D.E.A., E.N.S.I.E.G. Grenoble

Analyse cepstrale, visualisation sonographique et détection des formants. - Séminaire GALF-GRECO Analyse du signal de parole. Paris.

*. Contrat CNET; Caractéristiques individuelles

**. 15 locuteurs, 1 texte et 900 items chacun.



UN CHOIX D'EVENEMENTS POUR L'ORGANISATION TEMPORELLE DU SIGNAL DE PAROLE.

C. Abry, G. Benoît, L.J. Boë, R. Sock

I.C.P. Institut de Phonétique
Univ. des Langues et Lettres BP 25X 38100 Grenoble.

A set of events is proposed to reflect coordinative relations in the timing of the speech signal. 10 acoustic phonetic events prove to be sufficient in labelling the major discontinuities. An interpretative component provides 13 articulatory possibilities which can relate supraglottal stricture to main glottal activity in perspective of a coordinative grammar.

SEGMENTATION, REPRESENTATION ET PRODUCTION

On ne segmente que si l'on veut créer des segments (saucissonner le ver : échantillonner le signal) et/ou les retrouver (ver annelé : signal articulé).

Dire qu'on ne sait pas toujours retrouver dans le signal de parole les segments auxquels on s'attend - même lorsqu'on sait "ce qui a été prononcé" - est un truisme à modaliser selon les compétences du lecteur (l'"expert"), mais qui reste toujours plus ou moins vrai.

Segments en moins (élisions), segments en trop (épenthèses), segments insegmentables (amalgames), segments sursegmentables ("porte-manteaux") ; autant de "réalités" introduites par la complexité de projection des représentations phonétiques en réalisations, mais qui ne permettent pas pour autant de passer à une vision continue du signal *. Car des discontinuités il y en a **. On doit citer, bien sûr, à l'évidence, les plosions suivant le silence des occlusives sourdes. Et à l'autre extrême (les différentes phases d'une voyelle, par exemple), le repérage de ces discontinuités n'est plus immédiat.

A ce propos, la position que nous allons soutenir est que le nombre de discontinuités repérées sera d'autant plus grand :

1) qu'on se sera construit une représentation phonétique dont on puisse dériver le signal de parole ;

2) qu'on aura connaissance de sa production *. Le succès du repérage dépendra pour nous du degré d'adéquation de ces deux calculs mentaux.

Ce faisant nous n'oublions pas les risques que nous prenons : il n'y a actuellement ni grammaire du son toute faite, ni modèle articulatoire dynamique performant.

EVENEMENTS & COORDINATIONS vs. DISCONTINUITES ET SEGMENTS

Pas de segmentation sans transcription phonétique explicite ou implicite : la connaissance représentative permet de situer les phonèmes, de centrer les réalisations dans le signal, etc.

La connaissance productive permet, elle, de repérer davantage de discontinuités dans les phases de ces réalisations.

Un exemple : l'aboutissement de la closion (C) d'une occlusive sourde, avant la fin du voisement (VT : Voice Termination) et le silence, est une discontinuité (d'amplitude) qui n'est apparente sur le signal que pour celui qui la cherche (il y voit alors perte de structure formantique ou rebroussement de phase).

* Ce qu'on a tendance à faire trop souvent en opposant réalités phonologiques et phonétiques.

** Et à différentes échelles.

* Articulatori-acoustique. Un "filtre" ou une validation perceptif ne sont pas ici en cause.

En fait cette connaissance permet de modéliser le signal de parole comme un produit de "canaux" (2). Une discontinuité observée peut ainsi trouver son origine dans le "décrochement"/la mise en route d'au moins un de ces canaux : elle est donc le produit d'un ou de plusieurs événements, débuts d'onset ou fins d'offset sur les fonctions temporelles correspondantes.

Par exemple, cet événement qu'est l'apparition d'une périodicité due au voisement (VO : Voice Onset) peut coincider ou non avec l'apparition du bruit de plosion dû à la détente (R : Release) d'une occlusive, donnant un VOT (Voice Onset Time) nul, positif ou négatif.

Ainsi, alors qu'un segment minimal n'est qu'un contenu temporel (durée positive ou nulle) entre deux discontinuités successives, l'analyse de celles-ci en événements offre une réelle gamme de relations coordinatives (binaires ou n-aires) entre ceux de ces événements qui peuvent refléter les coordinations linguistiques apprises par le sujet parlant. On peut rappeller certaines de celles-ci comme la tenue des consonnes, le VOT (4), le VTT (Voice Termination Time (1), pour nous relation C-VT, soit voiced closure ou closure voicing in (4)), etc., sachant qu'il en reste plus d'une à construire ou à reformuler.

La prudence qui s'impose donc en la matière nous invite à utiliser la lecture en coordinations articulatoires du signal acoustique comme une heuristique et à ne pas faire dépendre trop étroitement les événements de cette grammaire à construire.

EVENEMENTS ACOUSTIQUES, PHONETIQUES ET ARTICULATOIRES

Toutes les discontinuités acoustiques repérées par le phonéticien ne sont pas en effet directement interprétables en événements articulatoires et certaines ne seront jamais que des événements acoustiques.

Un exemple pris encore sur le VOT : celui de LISKER et ABRAMSON (6) repose sur une coordination articulatoire (relation temporelle de VO par rapport à R), alors que le VOT défini par KLATT (5) ne reflète pas toujours une relation avec la source glottique (VO), mais avec la réponse acoustique du conduit vocal (apparition d'une structure formantique définie). Le "voisement vocalique" met en effet un certain temps à s'établir, même avec une attaque de voisement nette et dans un état "vocalique" (non obstruant) du tractus.

Ce VVO (Vocalic Voiced Onset : début vocalique voisé), ainsi que nous l'appelons, est par contre clairement interprétable comme une fin d'obstruction (R) du conduit vocal lorsqu'il suit la plosion d'une occlusive sonore. Quant au phénomène inverse - la perte de structure formantique "vocalique" ou VVT (Vocalic Voiced Termination : fin vocalique voisée) - on peut l'interpréter soit comme l'aboutissement d'une closion (C) - ce que nous avons déjà vu -, soit comme la fin d'une structure vocalique (en finale absolue).

La nécessité d'une composante d'interprétation productive distincte de la lecture des discontinuités acoustiques est donc de bonne méthode.

Mais il est un autre niveau que nous venons déjà d'utiliser implicitement, sinon sans méthode : celui de la connaissance représentative. La vocalicité que nous avons préfixée à VO (VVO) et VT (VVT) ne se satisfait pas d'une simple définition acoustique en formants, même en ajoutant l'absence de zéros dans le spectre (3). Cette restriction, qui pourrait convenir pour définir les discontinuités évidentes de séquences du type [ana, ala], ne suffira plus pour [amɛ̃, aɛ̃, etc.] - où l'on ne disputera pas sur la présence de zéros dans le spectre des voyelles nasales. Il faut donc recourir, dans l'état actuel de l'art, à des catégorisations phonétiques "vocalique" et "consonantique" de haut niveau, pour opérer un repérage des discontinuités satisfaisant tous ces cas.

Autant dire tout de suite, qu'on ne pourra pas non plus se contenter d'un étiquetage implicite du début du voisement des consonnes (CVO, Consonantal Voice Onset : début consonantique voisé) par la fin de celui des voyelles (VVT), et, respectivement, de la fin de ce voisement des consonnes (CVT, Consonantal Voice Termination : fin consonantique voisée) par VVO. Les séquences de type [am(e)na] obligent à cette relative redondance.

Mais il ne faudrait pourtant pas croire que cette catégorisation est "tout abstraite". Un exemple pris aux différentes phases des voyelle nasales françaises *, permettra de comprendre comment interviennent connaissance représentative et connaissance productive pour décider de ces attributs **.

Les six événements repérés jusqu'à présent sur les discontinuités l'ont été dans le domaine de la fonction de voisement (VO-VT).

* Cf. AUTESSERRE & al., dans ces mêmes J.E.P.

** Cf. ABRY & al., Propositions..., dans ces mêmes J.E.P., Fig.2.

Un autre "canal" du signal de parole que l'on pensera tout de suite à exploiter est la fonction de bruit (FO-FT, Frication Onset-Termination : début-fin de friction). Mais alors que l'interprétation articulatoire de voisement est univoquement d'origine glottique, la présence d'une excitation apériodique peut être d'origine supraglottique (plosions et frictions proprement dites) ou glottique (aspiration). Pour les premières, toute plosion contenant une part appréciable de friction, puisque le relâchement des articulateurs n'est pas instantanné, nous avons pris le parti de ne pas distinguer les deux * et nous parlerons indifféremment de plosion-friction, en étiquetant FO le début de plosion des occlusives comme le début de friction des constrictives bruissantes. En réalité la source glottique (aspiration) intervient aussi dans ce cas, puisqu'hormis certaines occlusives glottalisées (comme les éjectives, où la surpression orale est produite par l'élévation du larynx, glotte fermée), la plupart des plosions et frictions sont produites avec les deux sources. C'est surtout le cas des occlusives sourdes où la glotte est plus ou moins ouverte à l'instant de la plosion (à son maximum en anglais, proche du minimum en français) et aussi des constrictives bruissantes sourdes (où la glotte s'ouvre aussi puis se ferme, du début à la fin de la tenue). C'est le rapport des deux bruits, en faveur des sources supraglottiques, qui fait parler alors de plosion-friction et non de plosion-friction-aspiration, le bruit d'aspiration étant pratiquement "masqué" par les sources supérieures.

En prenant comme exemple la détente d'une occlusive : nous devrions en toute rigueur signaler, en même temps que le départ de la friction (FO), le début d'aspiration (Aspiration Onset : AO) ; puis étiqueter FT la fin de la plosion-friction d'origine supraglottique, qui s'arrête avant l'aspiration (Aspiration Termination : AT) ; laquelle se termine généralement à la reprise du voisement (VO) (quelquefois après quelques périodes bruitées, murmurées).

Mais nous savons qu'il est relativement difficile de toujours faire le partage entre les deux sources. On peut il est vrai, dans le cas des occlusives (parfois pour les constrictives bruissantes), utiliser la chute d'intensité du bruit, révélatrice de la faible énergie de la source glottique : on peut ainsi (avec un seuil) segmenter les deux phénomènes (plosion-friction-aspiration puis aspiration seule). Mais on rencontre – si l'on s'en tient à de telles décisions (AO-AT) – la plus grande complication étant donnée la différence de hauteur des voyelles avec lesquelles se co-produit la détente. On sait en effet que les voyelles dites hautes (i, y, u en français), lorsqu'elles sont chuchotées (souvent en fin d'énoncé ou après la détente des consonnes sourdes), présentent alors une source de bruit supraglottique. Comment donc décider sur le signal de la limite, dans les intermédiaires, entre une voyelle basse (de type a) chuchotée, pratiquement de source glottique, et une voyelle haute (de type i), à sources glottique et supraglottique ? Les début et fin de friction de l'une seraient-ils étiquetés AO-AT contre FO-FT pour l'autre ? De même, sachant qu'il existe une différence non négligeable entre la détente de [ki] en anglais et en français (le premier présentant un intervalle essentiellement aspiré [khi], après la plosion-friction de la consonne ; tandis qu'en français on observe une friction vocalique [kii] très marquée *), comment peut-on en rendre compte de manière non arbitraire à partir du signal ? Nous avons donc pensé que l'état de l'art nous permettait au mieux de confondre, dans une première lecture acoustique, toutes les sources apériodiques dans les étiquettes FO-FT, ceci pour ne pas provoquer d'emblée trop de décisions arbitraires.

Quant à l'appel à une catégorisation phonétique, nécessaire pour VO-VT, il l'est aussi pour FO-FT. Dans les cas de coproduction avec voyelle haute [kii], que nous venons de citer, la décision de repérer une discontinuité (en spectre, mais surtout en amplitude) après la plosion ne peut tenir à des arguments sur la différence des sources (présence de source apériodique supraglottique encore après cette fin de plosion), ni à des arguments sur la fonction de transfert (quand l'état vocalique, non obstruant, rendant possible le voisement, sera-t-il atteint après la plosion ?). C'est donc une décision catégorielle de haut niveau qui nous fera repérer CFT (Consonantal Frication Termination : fin consonantique fricative) dès la fin de la plosion (des plosions pour [k]) d'une occlusive ou, plus loin, la fin de friction d'une mi-occlusive [kçi]. Il en va de même pour repérer CFT dans la chute d'intensité caractéristique de la fin de tenue des constrictives bruissantes (nette avec les voyelles basses). Les débuts de friction (FO) appartenant à ces mêmes consonnes (début de plosion, début de friction) seront bien entendu étiquetés CFO (Consonantal Frication Onset : début consonantique fricatif).

Il ne nous semble par contre pas nécessaire d'avoir recours explicitement à une catégorisation phonétique VFO-VFT pour spécifier les parties dévoisées des voyelles, leur domaine étant aisément définissable par les autre événements.

LECTURE ARTICULATOIRE

Notre connaissance productive sera ici utilisée pour que les événements articulatoires reflètent les coordinations entre événements glottiques et supraglottiques.

Glottiquement VO et VT sont immédiatement interprétables (sans autres symboles).

Les événements supraglottiques seront conçus comme modulant différents degrés de striction du conduit vocal, atteints par les manoeuvres articulatoires tendant à la clusion (C) ou à la détente (R).

Pour prendre un exemple simple de transition voyelle-consonne en français, un VVT, fin de voyelle, correspondra à :

. une clusion maximale, de degré 4 (C4), quand il s'agira des occlusives [p,t,k,b,d,g] ;

une constriction de degré 3 (C3), pour les constrictives bruissantes [f,s,ʃ,ɛ,ʒ] ;

une constriction de degré 2 (C2), pour les constrictives non bruissantes ou approximantes (type fréquent du [v] français ;

. une occlusion de degré 1 (C1), pour les occlusives nasales [m,n] et la constrictive latérale [l].

Lorsqu'une discontinuité voyelle-nasale pourra être clairement lue, on pourra utiliser :

une occlusion de degré 0 (C0), en fermant le voile du palais pour passer d'une voyelle nasale (striction minimale de degré -1) à une voyelle orale [a,...,u] (striction 0, position phonatoire de référence, velum relevé) ou à une semi-voyelle [j,ɥ,w] *.

Dans le sens détente, nous aurons avec les mêmes degrés de striction :

R-1, en atteignant le niveau d'une voyelle nasale ;

. R0, pour les voyelles orales ;

. R1, pour les consonnes nasales et latérales ;

R2, pour les approximantes.

Pour la détente des occlusives bruissantes (type [t]) on passe, par R3 (CFO), niveau de plosion-friction, à R2 (CFT), niveau approximant, puis à R0, s'il s'agit d'une voyelle orale. On peut aussi passer directement à R2, si l'occlusive ne présente pas de bruit de plosion (cas parfois de [b]). La détente des constrictives bruissantes (CFT) passe elle aussi directement à R2 avant R0. Il arrive, avec les sourdes que, dans l'intervalle R2-VO, on ne sache pas bien quand commence l'état vocalique, non obstruant, R0. On peut alors par convention l'affecter à VVO.

. On peut enfin repérer sur le signal des détentes à bouche fermée, qui ne changent pas le niveau de striction, soit R4 (p. ex., la détente vélaire dans -g^{v}b-, qui permet une remontée caractéristique du fondamental).

Comme pour R0 après les consonnes sourdes, ce ne pourra être que par convention qu'on affectera "par excès" ou "par défaut" :

- en initiale d'énoncé :

. C4, aux occlusives sonores, VO n'étant pas synchrone de la clusion ;

. C0, aux attaques vocaliques orales, l'occlusion vélique s'étant produite préphonatoirement ;

- en finale :

R-1 à ces mêmes voyelles orales, lorsqu'il n'y a pas anticipation repérable acoustiquement de la position de repos.

PERSPECTIVES EN AVAL ET EN AMONT

Parvenus à ce point de notre inventaire événementiel, nous pouvons :

1) Segmenter le signal en utilisant toutes les discontinuités majeures, avec dix événements.

2) Structurer le signal en coordinations, en retenant les interprétations articulatoires de treize événements (C0,...,C4,R-1,...,R4 avec VO, VT).

3) Construire les classes phonétiques correspondant à ces coordinations.

Mais nous pouvons tout aussi bien, si cette approche ne nous convient pas :

1) Sursegmenter les segments entre nos événements *

2) Utiliser les segments directement pour constituer les classes phonétiques, sans passer par les coordinations.

3) Structurer des macro-segments, en regroupant autrement les événements que pour les coordinations.

Dans ces deux approches, où se situent l'aval et l'amont de nos

* Cf les problèmes posés par la classe des mi-occlusives.

* Se prolongeant parfois plus ou moins loin dans la voyelle, quand celle-ci n'est pas entièrement dévoisée (cas typique de "six" avec i bruissant [sis] ou chuchoté [sis]).

* Ces degrés de striction ont une vieille histoire au moins depuis le cours de SAUSSURE jusqu'à LADEFOGED (in (7)). On peut évidement en distinguer davantage. Nous ne retenons que les principales classes, celles qui nous semblent produire des discontinuités repérables sur le signal (ainsi confondons-nous voyelles et semi-voyelles).

* En phones homogènes ou cohérents cf. CAELEN & al. dans ces mêmes J.E.P.

événements ? Attendant une grammaire des coordinations temporelles, il nous est difficile de ne pas dire où vont nos préférences. Mais nous ne savons quelles pourront être les relations entre cette grammaire et celle des classes ? Y a-t-il forcément antécédence de celle-ci sur celle-là ? Autant de questions que nous n'allons pas prématurément poser. Nous défendons simplement l'option événementielle dans notre approche d'un signal, où l'importance de la dimension temporelle rend urgente l'intégration de cette dimension aux concepts mêmes de l'analyse ; c'est le cas pour l'événement ; c'est encore le cas pour l'organisation temporelle ou timing.

REMERCIEMENTS ET REDEVANCES

A toute l'équipe "Organisation temporelle" de l'Institut de Phonétique de Grenoble.

Aux équipes de Toulouse (CERFIA) et d'Aix-en-Provence (Phonétique) avec qui nous avons beaucoup échangé, ainsi qu'aux membres de la BD-son du GRECO Parole.

Aux "élèves" des cours CNET-ENST de Lannion qui ont subi notre première formulation d'ensemble.

A nos collègues grenoblois de l'I.C.P. qui nous ont patiemment écoutés.

Enfin, pour la relecture d'un premier manuscrit, à D. ABRY.

REFERENCES

(1) J.G. AGNELLO, "Measurements and analysis of visible speech", in "Measurements procedures in speech, hearing, and language", S.SINGH ed., Baltimore, London, Tokyo, pp. 379-397, 1975.

(2) C.G. FANT, "Speech sounds and features", Cambridge (Mass.), London, 1973.

(3) R. JAKOBSON, C.G. FANT & M. HALLE, "Preliminaries to speech analysis. The distinctive features and their correlates", Cambridge (Mass.), 1972.

(4) P. KEATING, W. LINKER & M. HUFFMAN, "Patterns in allophone distribution for voiced and voiceless stops", WPP 57, pp. 61-78, April 1983.

(5) D.H. KLATT, "Voice Onset Time, frication and aspiration in word-initial consonant clusters", J.S.H.R. 18, pp. 686-706, 1975.

(6) L. LISKER & A. ABRAMSON, "A cross language study of voicing in initial stops : acoustical measurements", Word 20, pp. 384-422, 1964.

(7) K. WILLIAMSON, "Multivalued features for consonants", Language 53, pp. 843-871, 1977

L'ORGANISATION TEMPORELLE CVCV CHEZ QUATRE LOCUTEURS FRANÇAIS OU FAUT-IL POURSUIVRE L'INVARIANT DANS LES VC ?

R. Sock, C. Benoît & C. Abry

I.C.P - Institut de Phonétique
Univ. des Langues et Lettres. BP 25X 38040 GRENOBLE.

In this paper we present few results obtained by temporal analysis of the speech signal. Mean durations of sounds as a function of context and position in two CVCV sequences spoken by four French speakers are here quantified and discussed. Furthermore, we study the problem of timing according to our conception of speech segments which is a reflection of gestual coordination. Finally, we point out the problem of "invariance" in VC sequences, resulting from the relative constancy of the consonant.

INTRODUCTION

Une hypothèse fondamentale en recherche sur la parole, veut que celle-ci soit caractérisée par une séquence de segments, chacun d'entre eux étant représenté par un faisceau de traits.

Retrouver l'équivalent de la constance perceptive dans le signal acoustique revient, soit à identifier, parmi les faisceaux, les propriétés invariantes qui caractérisent les unités linguistiques, et cela quelle que soit la transformation opérée sur le signal de parole, soit à repérer les relations "invariantes" entre événements articulatoires, à travers les variations de débit, d'accentuation, etc. (5).

A partir d'observations sur les variations de durées segmentales, nous tentons ici de retrouver une stabilité temporelle entre les gestes articulatoires dont les "cibles" ont été repérées sur le signal acoustique.

CORPUS ET PARAMETRES RETENUS

1. Nous utilisons un corpus de logatomes CVCV, conçu pour l'étude de combinaisons de sons maximisant et minimisant les effets de huit traits articulatoires.

2. Les logatomes sont insérés dans la phrase porteuse : C'est : "CVCV" ? Ça ?

3. Le corpus a été numérisé en chambre sourde au C.N.E.T. LANNION, avec un échantillonage à 16 KHz sur 16 bits. Les phrases apparaissaient en ordre aléatoire et à vitesse "normale" sur une console video.

4. Un éditeur de signal adapté (2) a permis l'étiquetage manuel des événements phonétiques (*).

5. Sur les 12 locuteurs enregistrés, 4 sont actuellement entièrement étiquetés, ce qui correspond à l'apposition de 13000 marqueurs sur le signal.

6. Nous n'avons retenu, dans une première approche, que 2 logatomes sur les 18 étiquetés en événements.

Ces deux logatomes, [baki] et [buki] nous permettront d'observer les bornes maximales et minimales du trait de voisement.

La production nous apprend que le prévoisement est favorisé par le [b], qui présente un volume plus grand que [d] ou [g] dans le conduit vocal. La flaccidité des joues semble être la grande responsable de cette optimisation du voisement.

De plus, le [u] présentant une meilleure coproduction aux lèvres que le [i], l'allongement du conduit vocal résultant favorise aussi le voisement.

Sur le plan acoustique, le contraste optimal de voisement est obtenu avec [b] et [k]. Si le premier l'optimise, le dernier présente un maximum de friction pendant la plosion, "dévoisant" ainsi la voyelle subséquente.

Les différences de durée intrinsèque entre [u] et [a] fournissent également le contraste maximal paradigmatique. Nous aurions pu établir ce contraste avec [i] et [a], mais la nature bruitée du [i] aurait nui à la sonorité du [b].

(*) Pour la définition des événements, voir ABRY & al. dans ces mêmes J.E.P.

7. Ici, nous n'avons utilisé que les 4 événements VO ("Voice Onset"), R ("Release" de degré 2), C ("Closure" de degré 4) et VT ("Voice Termination") suffisants à définir les quatre grands paramètres temporels (ou coordinations majeures) du signal : T1 ("tenue" de 'b' comprise entre VO et R), D.VOC1 (durée vocalique de a/u entre R et C), T2 ("tenue" de [k] entre C et R), D.VOC2 (durée vocalique de [i] entre R et VT) et DT (durée totale CVCV). Voir Fig. 1.

RESULTATS ET DISCUSSION

Les analyses statistiques nous ont permis de vérifier des résultats classiques, mais aussi de nous poser quelques questions sur la notion d'invariant, dans les coordinations entre unités constitutives.

Observons tout d'abord les caractéristiques intrinsèques et co-intrinsèques des segments : les valeurs moyennes (les 4 locuteurs confondus) des segments majeurs sont reportées dans le tableau ci-dessous.

Moyenne (ms)	[baki]	[buki]
T1	141	147
DVOC1	94	91
T2	132	116
DVOC2	234	247
DT	600	600

La figure 2 présente la structure temporelle réalisée par chacun des 4 locuteurs ainsi que leur "moyenne". En abscisse sont représentées les durées moyennes pour chaque segment majeur, tandis que l'ordonnée figure l'écart-type autour de cette "moyenne cible".

- la durée moyenne de [b], en contexte [-u], est très légèrement plus longue qu'en contexte [-a] chez 3 locuteurs (B,M,S).

- Comme l'ont déjà attesté certains auteurs, la voyelle ouverte [a] est plus longue que la [u] chez 2 locuteurs (B,D), l'inverse se manifestant cependant chez 1 locuteur (M)

- l'intervocalique [k] est plus longue en contexte [a-] qu'en contexte [u-] chez nos 4 locuteurs comme l'avaient déjà remarqué OSTRY et MUNHALL(4)Cette tendance s'expliquerait par la nature intrinsèque de la voyelle ouverte qui exige un abaissement de la mâchoire, augmentant ainsi la trajectoire de la consonne vélaire vers une cible avancée et beaucoup plus fermée la voyelle [i].

- la voyelle [i] subit un allongement de position finale accentuée (énoncé interrogatif) plus poussé pour [buki] que pour [baki] chez les 4 locuteurs. Par ailleurs, on vérifie que les deux [i], sous l'effet de cet accent syllabique, augmentant nettement en durée par rapport à leur valeur intrinsèque (la double de la valeur moyenne obtenue en position syllabique non finale chez ces mêmes locuteurs dans les logatomes [kiba] et [kibu]).

Ces résultats s'insèrent dans la conception du cycle gestuel vocalique, c-à-d la trajectoire d'une voyelle – à travers la consonne – vers une autre voyelle avec son influence sur la consonne. Ainsi, la durée de la consonne intervocalique est plus courte avec [u-] qu'avec [a-]. Par contre, le geste vocalique s'achève en allongeant le [i] final de [buki], ressemblant ainsi à une stratégie de compensation.

Ce cycle vocalique jouerait ainsi un rôle important dans la réorganisation temporelle prosodique de la syllabe.

Ces résultats nous ont ainsi permis d'établir un espace physique basé sur la durée phonétique des éléments constitutifs de nos logatomes. La quantification de cette dimension physiologique et acoustique sera élargie ultérieurement à d'autres réalisations adéquates pour explorer les bornes maximale et minimale des sons du français (1).

Le second axe d'intérêt de cette étude est de dégager l'organisation coordinative de nos locuteurs, en analysant les corrélations entre paramètres. Pour ce faire, nous nous sommes limités, sur le continuum du signal de parole, au champ VC (soit DT1 = DVOC1 + T2) et VCV (soit DT2 = DT1 + DVOC2) ; une décision due simplement au fait que la closion du [b] post-pausal n'est pas repérable. Le premier champ choisi nous rapproche d'ailleurs du paradigme expérimental adopté par TULLER & al.(5)en E.M.G. Leurs observations sur les invariants relationnels ont été vérifiés par la suite sur le plan kinésiologique. Nous essayerons ici de retrouver des résultats semblables au niveau acoustique en observant le "timing" des segments constitutifs.

Dans le champ VC (Fig. 3), nous constatons que la durée vocalique de V est corrélée positivement avec la durée totale de VC. Ceci est vrai, de manière générale, dans les 2 contextes et chez 3 locuteurs ; le plus rapide (S) manifestant une structure inverse (Tableau A).

Il est bon de signaler à ce point que les corrélations intra-locuteurs sont plus faibles que celles affichées tous locuteurs confondus : la confusion des locuteurs lents et rapides a en effet pour conséquence statistique, dans ce cas, leur réalignement sur la droite de régression.

Etant donné que le paramètre DT1 regroupe D.VOC1 et T2, il est logique, dans cette situation, de contrôler que la forte corrélation n'est pas due uniquement au fait d'inclure D.VOC1 dans DT1. Comme aucune corrélation n'est significative entre T2 et DT1, nous confirmons le fait, bien connu, que l'élasticité temporelle de la parole tient surtout au comportement de la voyelle (3), cette dernière représentant en quelque sorte le "noyau mou" de la syllabe, plus sensible à toute réduction ou expansion ; la tenue en représentant le noyau "dur", vu sa relative constance.

Explorons ensuite l'étendue du "timing" de nos locuteurs, en élargissant le champ d'observation de VC à VCV.

Quels sont les locuteurs qui maintiennent la même structuration temporelle que pour l'espace précédent ? En d'autres termes, est-ce que les quatre locuteurs présentant une corrélation positive entre D.VOC1 et la "nouvelle durée totale" DT2 ?

L'analyse (Tableau A) nous indique que cela n'est vrai que pour 2 locuteurs (B et D), et ceci dans les deux contextes vocaliques. Les deux autres locuteurs font, en quelque sorte, un "transfert" du "timing" sur la voyelle finale qui, elle, est fortement corrélée à DT2. Mais plus généralement, tous les locuteurs manifestent, dans les deux contextes, une corrélation assez élevée entre D.VOC2 et DT2. Nous pouvons, du point de vue de leur stratégie coordinative, séparer les locuteurs entre ceux qui limitent leur coordination au champ VC (M et S), et ceux qui opèrent sur les deux voyelles en relation avec DT2 (B et D). Par ailleurs, nous confirmons, ici aussi, l'augmentation des coefficients de corrélation lors du passage de l'intra-classe à l'inter-classe.

CONCLUSION

Cette étude nous a permis d'évoquer, à travers quatre locuteurs du français et leurs stratégies coordinatives, les délicats problèmes que pose l'organisation temporelle des segments. Mais faut-il – lorsqu'on trouve tous locuteurs confondus (ou toutes conditions) une étroite corrélation, voire un rapport constant dans un champ VC entre V et VC (en E.M.G., mouvement ou acoustique) – parler d'invariant sans rendre à C la constance qui lui appartient ?

(1) ABRY C. & BOE L.J. (1982)
Bull. Inst. Phon. Grenoble 10/11
(2) BENOIT C. (1984)
13e JEP GALF 211-213
(3) GAITENBY J. (1965) Haskins Labs. St. Rept. Speech Research 2, 1-12
(4) OSTRY D. & MUNHALL K. (1985)
J.A.S.A. 77, 640-648
(5) TULLER B. KELSO J.A.S. & HARRIS K.S. (1982) J. Exp. Psy. 8, 460-472

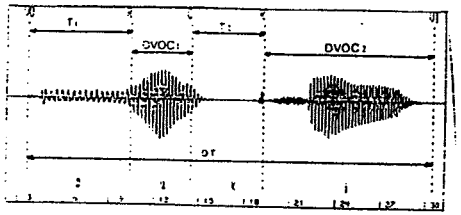


Figure 1

Logatome 'buki' par M.

DT : Durée totale du logatome

* Segments reflets de coordinations
T1 : Tenue consonantique de 'b'
DVOC1 : Durée vocalique de 'u'
T2 : Tenue consonantique de 'k'
DVOC2 : Durée vocalique de 'i'

* Evénements phonétiques :
VO : Début de voisement
R : Relâchement de degré 2
C : Closion de degré 4
VT : fin de voisement

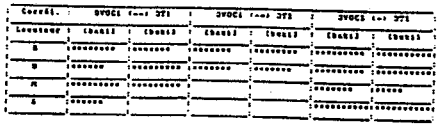


Tableau A

Corrélations entre paramètres pour 'baki' et 'buki' chez 4 locuteurs. Le nombre d'étoiles correspond à 10 fois le coefficient de corrélation.

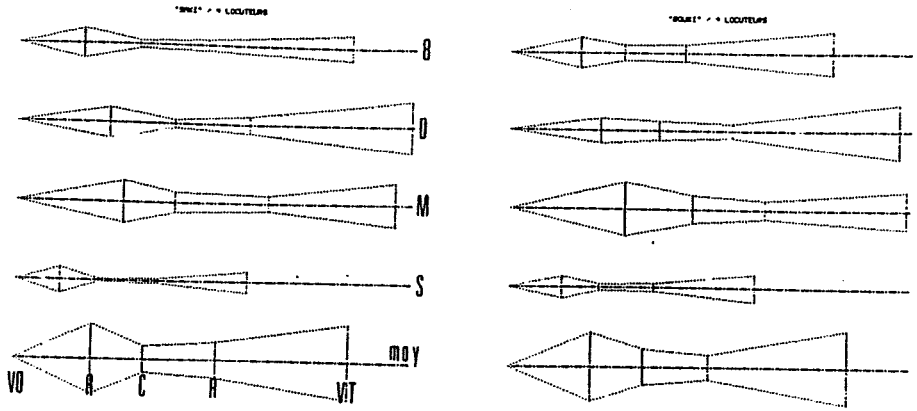


Figure 2

Abscisse : durées moyennes (max:710 ms)
Ordonnée : écart-types (max : 55 ms)

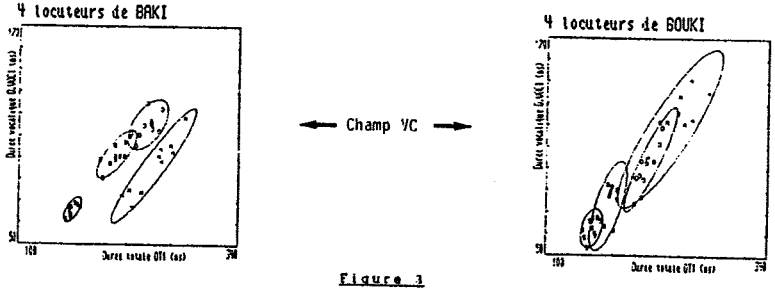
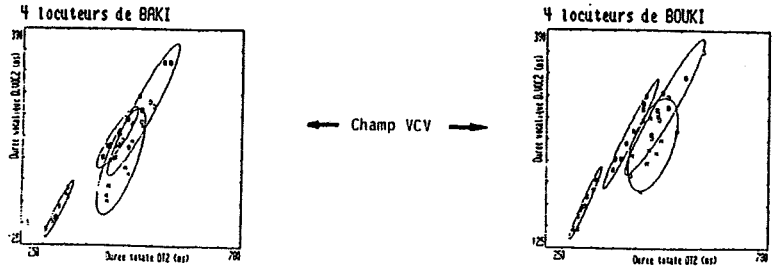


Figure 3

Ces projections représentent les ellipses de dispersion formées par les points de mesure de 2 paramètres temporels chez 4 locuteurs dans 2 conditions.



DE L'USAGE DES CORRELATIONS DANS L'ANALYSE SEGMENTALE DE LA PAROLE

C. Benoît

Centre National d'Etudes des Télécommunications, TSS/RCP, Route de Trégastel, 22301 Lannion, France.

ABSTRACT

A large number of investigators use the Pearson's product-moment correlation in analysing the temporal organization of speech. Some of their papers present results obtained by correlation between overlapping intervals. Our purpose is to show that exhibiting a high correlation could be a "statistical artifact" with such a method if no care is previously taken.

Therefore, the problem of part-whole correlations is mathematically examined in this paper.

INTRODUCTION

Parmi la littérature traitant de l'organisation temporelle de la parole, il est fréquent de trouver des analyses de données acoustiques, articulatoires, etc., réalisées au moyen des corrélations de Bravey-Pearson. Or, ces données temporelles consistent souvent en durées de segments dont l'un inclut l'autre.

Dans le domaine articulatoire, Bell-Berti et Harris [1], Tuller et al. [2,3], Tuller et Kelso [4,5], Harris et al. [6], confrontent la durée de la voyelle à celle du cycle voyelle-voyelle. Löfquist [7], puis Löfquist et Yoshioka [8,9] mesurent les corrélations obtenues entre la durée d'une partie d'occlusive (ou d'une fricative) et sa durée totale. En analyse E.M.G., Lubker [10] corrèle la durée d'une consonne avec celle d'un intervalle qui l'inclut. Soli [11] corrèle la durée d'une structure formantique stable avec elle-même augmentée de la durée de la transition qui la suit. Kohler [12] présente des coefficients de corrélation entre la syllabe et une partie de la syllabe. On rencontre également des corrélations entre segments "chevauchants" dans d'autres domaines, comme chez Folkins et Zimmerman [13] qui analysent des distances articulatoires incluses l'une dans l'autre.

Le premier, Barry [14] a noté que les fortes valeurs de coefficients obtenues par une méthode corrélant une partie avec le tout pouvait relever d'un "artefact statistique". En s'appuyant sur un raisonnement phonétique, il a suggéré l'emploi d'autres tests mais n'a pas mathématiquement prouvé ses objections.

C'est Munhall [15] qui, récemment, a été le plus loin dans une critique théorique de cette méthode en introduisant une formule de Cohen et Cohen [16] mettant en évidence l'existence d'un artefact statistique permettant de relativiser l'importance de telles corrélations.

Il restait à mener une étude mathématique théorique sur le problème simple de la corrélation entre un intervalle x (la partie) et un intervalle z = x+y (le tout)* afin de disséquer une méthode fréquemment utilisée par les chercheurs travaillant sur l'organisation temporelle de la parole.

* où x et y représentent le plus souvent deux variables temporelles contigues ; la voyelle et la consonne, p. ex.

FORMULES GENERALES DE CORRELATION

Soient n variables contigues x_i, i=1,n dont la somme est notée x_t. Nous noterons r_{ij} (resp. r_{it}) le coefficient de corrélation au sens de Bravey Pearson entre la variable x_i et la variable x_j (resp. la somme x_t).

On démontre facilement (voir Benoît, [17]) :

$$\sigma_t^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}\ \sigma_i\ \sigma_j \qquad (1)$$

et

$$\sigma_t = \sum_{i=1}^{n} r_{it} \cdot \sigma_i \qquad (2)$$

Ce qui signifie que l'écart-type du tout est le barycentre des n écarts-type des parties qui le constituent, pondérés par leurs coefficients de corrélation avec le tout. Ce coefficient est donc un INDICE de la contribution de la dispersion de chaque variable dans la dispersion de leur somme.

Mais il est plus intéressant de considérer une construction géométrique de cette relation à partir d'une représentation vectorielle des écarts-type : $\vec{\sigma}_t = \sigma_t \vec{u}_t$ et $\vec{\sigma}_i = \sigma_i \vec{u}_i$, i=1,n.

Cherchons les vecteurs $\vec{u}_i$ qui obéiraient à l'égalité :

$$\vec{\sigma}_t = \sum_{i=1}^{n} \vec{\sigma}_i = \sum_{i=1}^{n} \sigma_i \vec{u}_i \qquad (3)$$

Il suffit de projeter cette dernière relation sur l'axe défini par $\vec{u}_t$ pour obtenir :

$$\sigma_t = \sum_{i=1}^{n} \sigma_i \cdot \cos(\vec{u}_i, \vec{u}_t) \qquad (4)$$

Cette dernière relation, équivalente à la relation (2) montre que les coefficients de corrélation r_{ij} sont les cosinus directeurs de $\vec{\sigma}_t$ dans l'espace défini par les $\vec{\sigma}_i$.

L'angle entre $\vec{\sigma}_t$ et $\vec{\sigma}_i$ est donc égal à $\mathrm{Arccos}(r_{it})$

Par ailleurs, le calcul de la norme de $\vec{\sigma}_t$ donne :

$$\sigma_t^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_i\ \sigma_j \cos(\vec{u}_i, \vec{u}_j) \qquad (5)$$

Ce qui montre, par analogie avec (1) que les angles entre les dispersions de deux parties $\vec{\sigma}_i$ et $\vec{\sigma}_j$ valent $\mathrm{Arccos}(r_{ij})$.

Ces derniers résultats premettent alors de construire la représentation graphique des dispersions des variables et celle de leur somme en respectant leurs corrélations respectives. Voir la figure 1 dans le cas de deux variables.

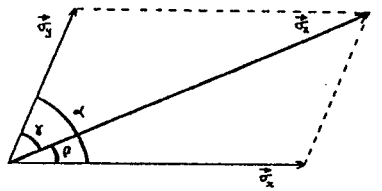


Fig. 1

Représentation géométrique des écarts-type de 3 variables x, y et z=x+y, de valeurs respectives $\sigma_x, \sigma_y, \sigma_z$. Les corrélations entre ces variables sont : $r_{xy} = \mathrm{Arccos}(\alpha)$, $r_{xz} = \mathrm{Arccos}(\beta)$, $r_{yz} = \mathrm{Arccos}(\gamma)$.

PROPRIETES DES CORRELATIONS ENTRE VARIABLES "CHEVAUCHANTES"

Nous nous limiterons ici au cas (le plus fréquent) de deux variables adjacentes x et y dont la somme vaut z.

Posons $r = r_{xy}$: corrélation entre les deux parties x et y,

$k = \sigma_x / \sigma_y$: rapport des dispersions des deux parties.

On démontre :

$$r_{xz} = (k + r) / (1 + 2kr + k^2)^{1/2} \qquad (6)$$

$$r_{yz} = (1 + kr) / (1 + 2kr + k^2)^{1/2} \qquad (7)^*$$

La comparaison de (6) et (7) permet d'établir que :

$$r_{xz} > r_{yz} \iff \sigma_x > \sigma_y \qquad (8)$$

La comparaison des coefficients de corrélation entre les parties et le tout se ramène donc à la comparaison des dispersions des parties.

* Notons que $r_{xz}(r,k) = r_{yz}(r,1/k)$

La figure 2 présente les domaines de variation des coefficients de corrélation r_{xz} et r_{yz} pour toutes les valeurs du rapport k quand r_{xy} varie de -1 à +1.

Ces deux domaines sont séparés par la courbe "d'isodispersion" définie par :

$$r_{xz} = r_{yz} = [(1 + r_{xy}) / 2]^{1/2}$$

On peut observer que, dans tous les cas, r_{xz} et r_{yz} sont simultanément supérieurs à r_{xy}.

Par ailleurs, une forte valeur positive de r_{xz} peut attester d'une valeur absolue élevée de la corrélation r_{xy} entre les parties et/ou d'une grande différence entre les dispersions.

Revenons ici à la formule définissant la "corrélation espérée" entre une partie et le tout, suggérée par Cohen et Cohen [16] et utilisée par Munhall [15] puis Löfquist [18] pour tester la validité de leurs corrélations observées :

$$r_{xz}^{o} = \sigma_x / (\sigma_x^2 + \sigma_y^2)^{1/2} \quad \text{(toujours > 0)}$$

obtenue en conservant aux deux variables x et y leurs dispersions observées mais en leur imposant une corrélation "espérée" $r_{xy}^{o} = 0$.

La figure 3 présente la construction géométrique obéissant à ce postulat simple.

Deux relations remarquables en découlent :

$$r_{xy} > 0 \quad (\Leftrightarrow) \quad r_{xz} > r_{xz}^{o} \text{ et } r_{yz} > r_{yz}^{o}$$

et :

$$r_{xz}^{o} / r_{yz}^{o} = \sigma_x / \sigma_y = k$$

Ce qui montre que l'étude de ce type de comparaison se ramène encore une fois à l'étude simple de la corrélation et du rapport des dispersions entre les parties.

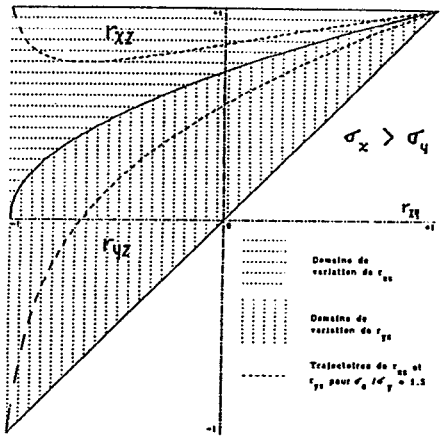


Fig. 2
Domaines de variation des corrélations de chacune des variables x et y avec leur somme z=x+y, en fonction du coefficient r_{xy}, obtenus en faisant varier le rapport $k = \sigma_x / \sigma_y$ de 1 (courbe de séparation en traits pleins) à l'infini (droite sup. horizontale pour r_{xz} et inf. oblique pour r_{yz}).

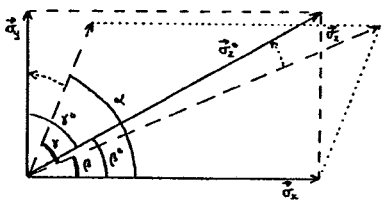


Fig. 3
Représentation géométrique analogue à la Fig. 1 pour les traits tiretés (corrélations observées : α, β et γ) et construction en traits pleins obtenue à partir des corrélations "espérées" : $\alpha^o = 90^o$, β^o et γ^o telles que

$$r_{xy}^{o} = 0, \; r_{xz}^{o} = \text{Arccost}(\beta^o), \; r_{yz} = \text{Arccos}(\beta^o)$$

CONCLUSION

Cette brève étude mathématique ne préjuge pas de la signification des corrélations calculées : celle-ci dépend de la méthode statistique utilisée et du nombre d'items constituant les variables.

Nous ne prétendons pas critiquer les conclusions des auteurs ayant utilisé ce type de corrélations, car elles ne dépendent pas forcément de la manière dont sont présentés leurs résultats.

Notre but est ici de contribuer à une meilleure maîtrise des corrélations entre segments "chevauchants" chez les chercheurs étudiant l'organisation temporelle de la parole.

Ainsi, comme Barry [14] l'a suggéré dans le cas de deux variables, nous préférons raisonner en termes de corrélations entre les parties plutôt qu'entre la partie et le tout, aussi artificielles que puissent paraître les coordinations articulatoires entre segments contigus.

De même, et comme le pense Munhall, il est plus simple et sans doute plus "parlant" d'analyser les résultats par les dispersions relatives des parties plutôt que par leurs corrélations avec le tout : cf. la formule (8).

Toutes les recherches d'une relation de phase invariante dans les études sur le "timing" articulatoire (de [1] à [9]) n'ont en réalité jamais reposé que sur cette comparaison entre les dispersions des deux parties adjacentes, quelles que furent les transformations mathématiques ou les déformations de la réalité utilisées.

Enfin, ce type de corrélations a néanmoins pu être utilisée avec profit (Benoit, [17]) dans l'analyse de plus de deux parties dont on veut connaître les relations avec le tout, dès lors qu'elles ne sont perçues que comme un indice de la contribution de leurs dispersions dans celle de leur somme : cf. la formule (2).

REFERENCES

[1] Bell-Berti, F. and Harris, K.S. (1981). "A temporal model of speech production," Phonetica 38, 9-20.

[2] Tuller, B., Kelso, J.A.S. and Harris, K.S. (1981). "Phase relationships among articulator muscles as a function of rate and stress," Hask. Lab. St. Rep. SR-65, 63-90.

[3] Tuller, B., Kelso, J.A.S. and Harris, K.S. (1982). "Interarticulator phasing as an index of temporal regularity in speech," J. Exp. Psychol. : H.P.P. 8, 460-472.

[4] Tuller, B. and Kelso, J.A.S. (1983). "Converging evidences for the role of relative timing in speech : a reply to Barry," J. Exp. Psychol. : H.P.P. 9, 829-833.

[5] Tuller, B and Kelso, J.A.S. (1984). "The timing of articulatory gestures : Evidence for relational invariants,"J. Acoust. Soc. Am. 76 1030-1036.

[6] Harris, K.S , Tuller, B. and Kelso, J.A.S. (in press). "Temporal invariance in the production of speech," in Invariance and Variability of Speech Processes, edited by J. Perkel (Eribaum, Hillsdale, NJ).

[7] Löfquist, A. (1980). "Interarticulator programming in stop production," J. Phon. 8, 475-490.

[8] Löfquist, A. and Yoshioka, H. (1981). "Interarticulator programming in obstruent production," Phonetica 38, 21-34.

[9] Löfquist, A. and Yoshioka, H. (1984). "Intrasegmental timing : laryngeal-oral coordination in voiceless consonant production," Speech Com. 3, 279-289.

[10] Lubker, J. (1981). "Temporal aspects of speech production : anticipatory labial coarticulation," Phonetica 38, 51-65.

[11] Soli, S.D. (1982). "Structure and duration of vowels together specify fricative voicing," J. Acoust. Soc. Am. 72, 366-378.

[12] Kohler, K.J. (in press). "Invariance and Variability in Speech Timing : Utterance to segment in German," in Invariance and Variability of Speech Processes, edited by J. Perkel (Eribaum, Hillsdale, NJ).

[13] Folkins, J.W. and Zimmermann, G.N. (1982). "Lip and jaw interaction during speech : Responses to perturbation of lower lip movement prior to bilabial closures," J. Acoust. Soc. Am. 67, 271-275.

[14] Barry, W.J. (1983). "Some problems of interarticulator phasing as an index of temporal regularity in speech," J. Exp. Psychol. : H.P.P. 9, 826-828.

[15] Munhall, K. (1985). "An examination of intra-articulator relative timing," J. Acoust. Soc. Am. 78, 1548-1553.

[16] Cohen, J. and Cohen, P. (1975). Applied Multiple Regression/Correlation, Analysis for the Behavioral Sciences (Eribaum, Hillsdale, NJ).

[17] Benoit, C. (1985). "Caractéristiques individuelles du locuteur par analyse des données temporelles du signal de parole," Thèse de Docteur-Ingenieur : E.N.S.E.R. Grenoble.

[18] Löfquist, A. (in press). "Stability and Change : Comments on Kelso, Saltzman and Tuller," J. Phon.

REMERCIEMENTS

à C. Abry pour ses suggestions, à S. Maeda pour ses remarques et à K. Munhal pour ses encouragements.

VOTS ET VTT EN FRANCAIS

R. Sock & C. Benoit

Institut de la Communication Parlée - Institut de Phonétique - Grenoble
Département R.C.P. - CNET Lannion

ABSTRACT

Voice Termination Time and two specifications for VOT (LISKER & ABRAMSON vs. KLATT) are measured for 5 French speakers producing CVCV logatemes in a carrier sentence. Results indicate that vowel height enhances maximal contrasts for both VTT and VOT. As concerns minimal contrasts, negative vs. positive values prove to be better discriminators than merely positive ones (KLATT). VTT is also a separator. Its relevance in intrasegmental timing is finally investigated in relation to the invariance issue.

La plupart des recherches en acoustique portant sur le voisement se sont très souvent concentrées sur le VOT [1 & 2] en tant qu'indice principal de ce trait, comme lié à son timing par rapport au geste supraglottique de détente - à l'initiale et à l'intervocalique. Mais il est tout aussi loisible d'examiner le comportement d'un autre indice, lié lui aussi au timing du geste glottique par rapport à la closion supraglottique : le VTT [3]. Ce délai d'arrêt du voisement - acoustiquement présent, lui, à l'intervocalique et à la finale - est déjà intervenu dans la spécification du trait (comme voiced closure ou closure voicing in [4]), en ce qui concerne les préférences distributionnelles, et les conséquences de ces préférences pour les changements phonétiques. Il peut être relié enfin plus généralement, tout comme le VOT, aussi aux contraintes d'ensemble régissant le timing du geste glottique [5].

On tentera d'abord ici - en explorant les bornes du voisement, avec le contraste /b/ vs. /k/ [6] - d'établir les espaces maximaux [7] de ces deux indices.

Pour le VOT dans sa première définition, celle de LISKER et ABRAMSON, on pourra comparer ces résultats extrêmes avec ceux obtenus pour le français par nos prédécesseurs (de [8] à [9] en passant par [10]). La démonstration de l'efficacité perceptive de ces VOT négatifs et positifs en français n'est plus à faire depuis les travaux de [11].

Nous ajouterons à ces mesures traditionnelles du VOT celles effectuées selon la conception de KLATT, pour tester la rentabilité discriminante de cette mesure en français. En effet, si son utilisation ne pose pas de problèmes en anglais (et/ou en suédois, par exemple [4]), où les consonnes non voisées présentent de longs délais aspirés, son efficacité discriminante en français peut être mise en doute, non pas à cause des problèmes de détection perceptive qu'il pourrait poser (au contraire [12]), mais tout simplement parce que le chevauchement des distributions observées semble trop important.

Nous ferons de même pour le VTT qui, lui, n'a jamais, à notre connaissance, été étudié directement en français. Une fois précisées ses bornes nous exposerons sa contribution objective à la discrimination statistique du voisement.

A notre avis l'efficacité perceptive de cet indice n'a été démontrée qu'indirectement : les manipulations sur la réduction de la tenue des consonnes occlusives non voisées [13] peuvent être un argument en ce sens. Plus généralement, l'étude de l'organisation temporelle du VTT doit être envisagée dans la perspective du contrôle du geste glottique par rapport aux deux commandes consonantiques de closion et de détente qui produisent cette tenue. Replacée dans le contexte des recherches actuelles sur le phasage (phasing) de ces commandes en tant que révélateur d'un éventuel programme moteur intersegmental [14], cette étude doit permettre de comparer les manifestations acoustiques de notre VTT aux résultats obtenus par [5], en ce qui concerne la production d'une coordination intrasegmentale.

Nous utiliserons dans ce travail les variations interlocuteurs comme un paradigme parmi d'autres (comme les variations de vitesse d'articulation, d'accent, etc.), pour mettre en évidence les résistivités, sinon les invariants, dans l'organisation temporelle du voisement.

CORPUS ET MESURES

Nous utilisons un corpus numérisé de logatomes CVCV, conçu pour l'étude de sons maximisant les effets de huit traits articulatoires du français [15].

Ces 18 logatomes, CVCV - insérés dans la phrase porteuse C'est : "CVCV" ? Ça ? - ont été prononcés (10 réalisations de chaque) par 12 locuteurs. Nous ne présenterons ici graphiquement que les résultats obtenus sur 5 d'entre eux, illustrés par 6 logatomes, particulièrement choisis pour rendre compte du voisement dans les occlusives.

Un éditeur de signal adapté [16] a permis la segmentation manuelle en événements acoustiques [17].

Quatre paramètres ont été retenus : VOT, VTT, D.VOC et TENUE, reflétant les coordinations glottiques et supraglottiques [17] (cf. APPENDICE, infra).

Les VOTs

1. Le contraste maximal /bu/ vs. /ki/

Nous remarquerons que pour le VOT de LISKER & ABRAMSON le contraste maximal correspond à notre hypothèse de départ. A l'initiale nous avions posé que le prévoisement du /b/ dans /buki/ présenterait le VOT négatif maximum (-137 ms.); ceci tient à des cavités (pharyngale et buccale) assez larges par rapport aux autres occlusives voisées, à une flaccidité des joues propre à la production du /b/ [18] et de surcroît, à sa coproduction avec /u/ qui favorise ainsi le voisement : le volume propre à la voyelle est plus important que celui d'un /a/ par exemple; ce que nous confirmons avec /baki/ (-126 ms.). Il faut noter tout de même que la différence observée entre les deux prévoisements est minime (11 ms.) et qu'à cause des fortes dispersions elle n'est pas significative tous locuteurs confondus. Ce n'est que ceteris paribus (à même contexte et même locuteur) que l'on peut observer des résultats significatifs sur de tels phénomènes cointrinsèques [19]. Pour le VOT positif on obtient l'autre borne avec /kiba/ ou avec /kibu/ (53 et 49 ms.), proches de /kupa/ (45 ms.), bien loin de /kapa/ (26 ms.).

Les bornes du VOT s'établissent donc par le contraste /bu/ vs. /ki/ (Fig. 1). L'histogramme nous indique que le VOT pour /k/ possède une cible plus précise absolument que celle pour /b/, si l'on s'en tient à la dispersion (les écarts-types) autour des valeurs modales. Mais notons que les cibles sont relativement aussi précises l'une que l'autre (coefficients de variation : 34 % et 33 %).

Avec le VOT défini par KLATT, nos hypothèses de départ restent les mêmes. Nous retrouvons à l'initiale le contraste maximal avec /buki/ (7 ms., significativement différent de /baki/, 8 ms.) contre /kiba/ ou /kibu/ (63 et 60 ms.), proches de /kupa/ (53 ms.), significativement différents de /kapa/ (34 ms.) (cf. Fig. 2). Cette fois-ci c'est le /k/ qui présente les réalisations les moins dispersées relativement (27 % contre 43 %).

Ces items (kibu vs. buki) servent aussi à établir les bornes du VOT à l'intervocalique avec des résultats comparables.

2. L'opposition /ba/ vs. /pa/

Il ne nous est pas difficile d'opposer à l'initiale /b/ à /p/ par le VOT de LISKER & ABRAMSON (avec -126 ms. pour /baki/ contre 7 ms. pour /paki/). Notons que sur une cible extrêmement plus courte que celle du /b/, on observe pour le /p/ (Fig. 3) une imprécision relativement forte (coefficients de variation : 57 % contre 37 %).

Cette opposition par le VOT devient moins praticable en adoptant pour le français la conception de KLATT. Il est intéressant de noter tout de même que les valeurs moyennes (cf. les modes sur Fig. 4) restent assez contrastées (8 ms. pour /b/ contre 14 ms. pour /p/) et que la différence reste statistiquement très significative (t = 7.63; seuil à 90 % = 1.63). Dans ce cas notons que c'est le VOT du /b/ qui possède une cible relativement moins précise que celle du /p/ (38 % contre 29 %).

Le VTT

1. Le contraste maximal /ak/ vs. /ib/

Le VTT minimum, comme nous l'avions prévu, est obtenu avec /ak/ (/paki/, /baki/ ou /paku/, environ 30 ms., contre /kupa/, 38 ms., comme /buki/). C'est la coproduction qui présente la plus petite cavité pharyngale (la voyelle /a/ avec une occlusion vélaire), conditions préjudiciables au prolongement du voisement; le rétrécissement pharyngal de /a/ semble d'ailleurs suffire, à en juger par /kapa/ (30 ms. aussi). Le VTT est à son maximum, bien entendu, quand il est égal au VOT négatif (99 ms. pour /kibu/ (Fig. 5), c'est-à-dire lorsque le prévoisement ne s'arrête pas avant la détente comme cela se produit parfois pour /b/.

Les valeurs moyennes indiquent qu'il y aurait à première vue un lien entre le VTT et la qualité - quantité de la voyelle (mais sans qu'il y ait pour autant une corrélation systématique entre la durée de celle-ci et le VTT). On peut dire simplement, d'après ces résultats, que les voyelles hautes présentent un VTT significativement plus important que les basses. C'est un phénomène cointrinsèque d'une autre origine que celui que nous avions pu constater ailleurs [20] pour les voyelles phonologiquement longues (à VTT long) opposées aux brèves (à VTT bref), quel que soit l'effet de hauteur de la voyelle.

2. L'opposition /ba/ vs. /pa/

Ne disposant pas dans notre corpus d'une paire minimale à l'intervocalique, nous utilisons pour révéler cet indice la quasi-paire /kupa/ vs. /kiba/, /pa/ et /ba/ étant tous deux précédés ici d'une voyelle haute (à même effet cointrinsèque). Les deux populations restent nettement séparées, avec des distributions comparables (Fig. 6). Relativement, ce sont les réalisations du VTT avec /p/ qui affichent un comportement plus dispersé (34 % contre 13 %).

3. Le timing intrasegmental du VTT

Si l'organisation temporelle intrasegmentale du VOT est relativement bien connue (pour le français voir en particulier [21], où burst ≙ VOT de

KLATT), on ne possède pas autant d'informations sur la place que prend le VTT dans les coordinations consonantiques. Ainsi pour les occlusives françaises non voisées, l'intervalle de silence précédant la détente a été sans doute le plus mesuré [8]. Mais il ne permet évidemment pas de relier directement le voisement à la closion et donc à la tenue.

On constate entre la tenue et le VTT une corrélation négative : plus la tenue du locuteur est longue plus le VTT sera court. C'est une tendance générale qui n'est contredite que par un seul de nos locuteurs.

Sachant qu'il a été réaffirmé récemment par LÖFQVIST que la phase d'ouverture de la glotte restait relativement invariante par rapport à la tenue, quelles que soient les variations de vitesse d'articulation [5], nous avons examiné le VTT en pourcentage des variations de la tenue. On remarque (Fig. 7) que l'allongement de cette dernière a tendance à laisser le VTT proportionnellement inchangé au delà d'un certain "seuil", la pente de régression des réalisations de plus grande densité étant moins prononcée que celle des occurrences, plus rares, qui présentent des tenues courtes. Le pourcentage de VTT tend à augmenter rapidement avec la réduction de la tenue et,bien que l'on soit ici encore loin du voisement total (avec au maximum 30 % de VTT pour environ 110 ms de tenue),la dérive s'annonce rapide vers les occlusives affaiblies (flapped), pour lesquelles l'indice de prévoisement ne peut plus jouer dans la catégorisation du trait.

En essayant de relier l'hypothèse de LÖFQVIST sur la production du geste glottique - replacée récemment dans le cadre de la Théorie de l'Action - au signal acoustique, on assisterait d'après nos données au même phénomène pour les coordinations intrasegmentales que dans le domaine intersegmental, à savoir : les variations d'amplitude de l'ouverture glottique en fonction de la durée de la tenue (comme les variations d'amplitude de la cible vocalique en fonction de la latence consonantique, [14]) laisseraient relativement invariantes les phases articulatoires; alors que pour le signal acoustique, les différentes contraintes aérodynamiques permettraient plus difficilement de retrouver ces constances de phases . (Pour une discussion sur cette invariance relative articulatoire (?) et/ou acoustique (?),cf. [22]). Reste à savoir,dans ces conditions,sur quelles constances perceptives peut reposer la récupération des indices du voisement [23].

Remerciements à :

C. ABRY pour ses commentaires et le temps qu'il a consacré à la réalisation de ce travail; D. VUILLET pour la frappe du texte et pour ses conseils d'ordre pratique; A. SANCHEZ pour son soutien informatique.

REFERENCES

[1] L. LISKER & A.S. ABRAMSON, Word 20, 384-422, 1964.
[2] D.H. KLATT, J. Speech Hearing Res. 18, 686-706, 1975.
[3] J.G. AGNELLO, in : Measurement Procedures in Speech, Hearing, and Language. S. SINGH (Ed.), Univ. Park Press, 379-397, 1975.
[4] P. KEATING, UCLA-WPP 57, 26-60, 1983.
[5] A. LÖFQVIST & H. YOSHIOKA, Phonetica 38, 21-34, 1981.
[6] K.N. STEVENS, Symposium 1, 9th Int. Congr. Phon. Sci. 3, 190, 1979.
[7] C. ABRY & L.J. BOË, Bull. Inst. Phonét. Grenoble 10/11, 1-12, 1981-1982.
[8] W. SERNICLAES, 13èmes JEP du GCP-GALF, 69-78, 1984.
[9] J.P. GOUDAILLIER, 14èmes JEP du GCP-GALF, 47-50, 1985.
[10] D. ROSTOLLAND, C. PARANT, A. TAKAHASHI & E. PANDALES, 14èmes JEP du GCP-GALF, 179-182, 1985.
[11] W. SERNICLAES & P. BEJSTER, R.A. de l'Inst. Phonét. de Bruxelles 10/1, 83-94, 1978.
[12] B. DELGUTTE, Thèse d'Etat, Paris 6, 1984.
[13] J. SWEERTS & M. WAJSKOP, R.A. de l'Inst. Phonét. de Bruxelles 7/1, 35-47, 1973.
[14] C. FOWLER, P. RUBIN, R. REMEZ & M. TURVEY, in Language Production, B. BUTTERWORTH (Ed.), 373-420, 1979.
[15] C. BENOIT, Thèse Doct.Ing., INP Grenoble, 1985.
[16] C. BENOIT, 13èmes JEP du GCP-GALF, 211-213, 1984.
[17] C. ABRY, C. BENOIT & R. SOCK, Docum. GRECO Comm. Parlée, 1985.
[18] B. SMITH & J. WESTBURY, Paper presented at the 89 th Meeting of A.S.A., 1975.
[19] A. DI CRISTO, Thèse d'Etat, Univ. de Provence, 1978.
[20] R. SOCK, Thèse Doct. 3ème Cycle, Inst. Phon. de Grenoble, 1983.
[21] M. WAJSKOP, in : Frontiers of Speech Comm. Res., LINDBLOM & ÖHMAN (Eds.), Academic Press, 109-123, 1979.
[22] C. BENOIT & C. ABRY, 12e ICA, 1986.
[23] Q. SUMMERFIELD, J. of Exp . Psychol. 7/5, 1074-1095, 1981.

APPENDICE

1. Le VOT de LISKER & ABRAMSON appliqué au français est :
   . positif pour les occlusives non voisées,du début de la plosion-friction (événement CFO in [17]) au début du voisement (VO);
   . négatif pour les occlusives voisées :
   -à l'initiale, du début du voisement (VO) à CFO;
   -à l'intervocalique, de la closion (événement acoustique VVT in [17], correspondant à une perte de structure formantique vocalique) à CFO.
2. Le VOT de KLATT est toujours positif : de CFO à VVO (établissement de la structure formantique vocalique de la voyelle suivant la plosive).
3. Le VTT va de la closion (VVT) à la fin du voisement (VT) pendant la phase d'occlusion (il est ≤ au VOT négatif de LISKER & ABRAMSON).
4. La TENUE de la consonne va,pour nous,de la closion (VVT) à la fin de la plosion-friction proprement consonantique (CFT); ce qui harmonise sur ce point plosives, fricatives et affriquées.
5. D.VOC., état vocalique du conduit,occupe le reste du cycle détente-détente (CFT-CFT) : soit de CFT à VVT.

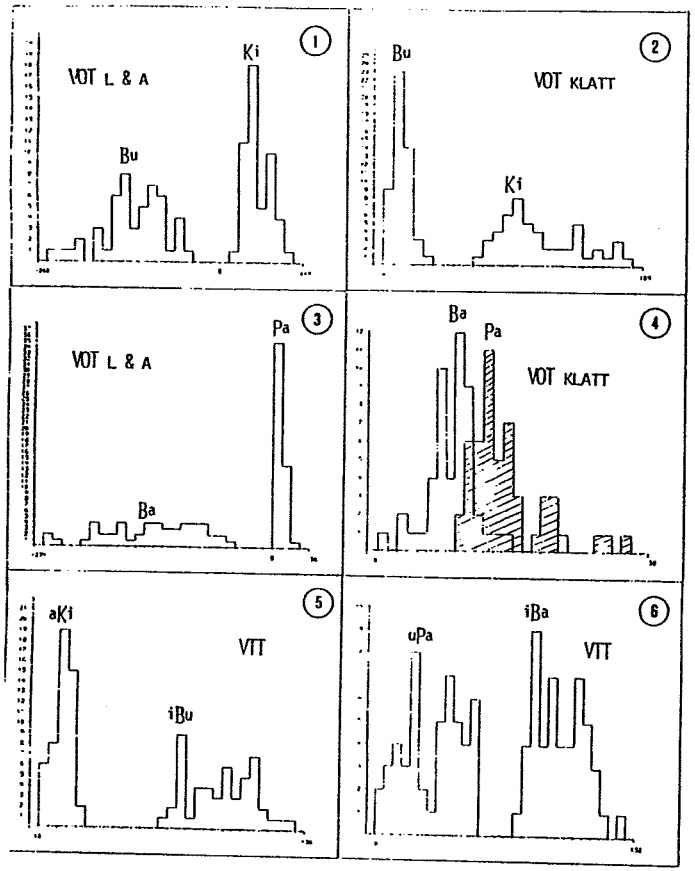


Fig. 1-2. Contraste maximal, VOT LISKER & ABRAMSON vs. VOT KLATT pour /buki/ et /kibu/ (5 loc. x 10).

Fig. 3-4. Opposition minimale. VOT LISKER & ABRAMSON vs. VOT KLATT pour /baki/ et /paki/ (5 loc. x 10).

Fig. 5. Contraste maximal. VTT de /paki/ et /kibu/ (5 loc. x 10).

Fig. 6. Opposition minimale. VTT de /kupa/ et /kiba/ (5 loc. x 10).

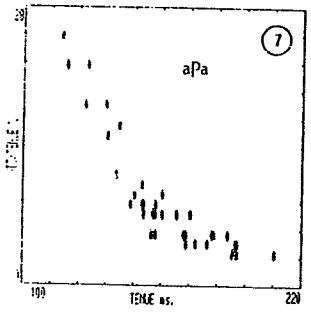


Fig. 7. Le phasage du VTT par rapport à la TENUE pour /kapa/ (4 loc. x 10).

DE LA PRODUCTION A L'EXTRACTION, L'ETAT D'UN CHANTIER.

F. EMERARD, C. BENOIT.

Centre National d'Etude des Télécommunications
B.P. 40
22301 LANNION Cedex

A prosodic data base has been built from the analysis of two corpa read by a male and a female speakers native of French. It corresponds to sixteen minutes of labelled speech segments.
For each segment (a phoneme, a VOT or a pause), a special coding includes syntactic, lexical, phonetic informations and their durations. Three fundamental frequencies values are also added for the vowels.
A systematic investigation of this database allows the constitution of a knowledge base. It will be applied for improving the modelling of French prosody in the actual CNET text-to-speech synthesis in order to make it more natural sounding.

I INTRODUCTION

Aujourd'hui comme en 1980 et comme le soulignait déjà LIENARD [1], la synthèse à partir du texte n'est toujours pas un problème résolu... Problèmes d'analyse du texte à synthétiser, de concaténation des éléments choisis, d'analyse du signal, de synthétiseur, de naturel de la voix, d'etc.

Concernant le traitement prosodique pour la synthèse, l'amélioration - sans faire appel à des notions sémantiques - peut passer à la fois par une analyse et une synthèse plus fines des phénomènes rythmiques et mélodiques et par une synthèse plus "variée" de ces phénomènes.

C'est à cette fin qu'une Base de Données a été spécialement constituée, organisée autour de l'étiquetage en phones d'un texte, en prenant en compte dans son codage les possibilités d'interrogation (par mots-clés) pour l'analyse et la synthèse de la durée segmentale et syllabique, du fondamental et des pauses.

II CONTENU DE LA BASE

Plutôt que d'énoncer des listes de résultats encore parcellaires (pour deux locuteurs, durée des consonnes, durée et répartition syntaxique des pauses, etc.), il nous paraît plus intéressant de préciser ici - pourquoi pas au titre de la contribution aux travaux BD-Sons du GRECO ? - une méthode de constitution et d'interrogation d'une base de données prosodiques. Ainsi peut-être cette méthode pourra-t-elle s'enrichir de propositions pour un codage et des interrogations différentes, s'élargir à l'étiquetage d'autres langues et faciliter par là-même des études prosodiques comparatives *.

Cette base de données est constituée d'un fichier ASCII par locuteur (500 Koctets chacun) et interrogeable sur un LSI 11/73. Elle contient aujourd'hui les données concernant 150 phrases énonciatives lues par deux locuteurs français, ce qui permet, pour chacun des locuteurs, l'analyse d'environ huit minutes de parole, de quelque 6000 segments acoustiques, 200 pauses, 2300 syllabes et voyelles et plus de 1500 mots.

L'étiquetage a été réalisé en deux étapes successives, l'une manuelle, l'autre automatique.

1) Informations acquises manuellement :

- un étiquetage du corpus en code phonétique "ad hoc" parmi les 38 types de segments (deux caractères par type) - dont la pause interne et la pause de fin de phrase - répertoriés pour le français a été introduit ligne par ligne sous éditeur,
- une valeur de durée de chaque segment, introduite au moyen d'une tablette d'acquisition graphique et d'une souris (programme adapté de l'Institut de Phonétique de Grenoble), à partir de spectrogrammes [2] segmentés manuellement,
- trois valeurs de Fo pour chaque voyelle, obtenues suivant le même principe,
- un codage écrit sous éditeur en mot grammatical (MG) ou mot lexical (ML) sur chaque ligne correspondant à un début de mot,
- un codage de la position de chaque syllabe pour chaque mot ; par exemple 2-4 marque le premier segment de la deuxième syllabe d'un mot de quatre syllabes,
- un codage catégoriel pour chaque mot : 27 "positions de mot" ont été nécessaires (x 2 selon que "suivi de pause" ou non) ; dernier mot de phrase, mot situé en fin d'incise, dernier mot avant le verbe, p. ex.

Notons que les deux locuteurs ayant prononcé le même corpus, les deux fichiers sont isomorphes pour le codage "manuscrit", à quelques rares différences près sur les "e muets", les pauses et les VOT. Ainsi, la transcription de l'un a pu être dupliquée sur l'autre automatiquement.

2) Informations acquises automatiquement :

- un codage de la position de chacun des segments dans chaque syllabe : l'étiquette 2/5 repérant le deuxième segment d'une syllabe en comportant 5,

* Une base constituée sur le même modèle et pour un corpus le plus proche possible lexicalement et syntaxiquement est en cours pour le portugais : VIANA, à paraître

- un codage de la nature de chaque syllabe, ouverte (a) ou fermée (f),
- un codage par "classe phonétique" sur chaque ligne : pv(nv)vo signale que le segment courant est une voyelle nasale précédée d'une plosive voisée et suivie d'une voyelle orale, par exemple,
- un codage phonétique de la syllabe, avec indication de sa durée, sur chaque phone.

La figure 1 présente le codage final d'une phrase de ce corpus obtenu après application de 10 programmes intermédiaires. Ce fichier est alors prêt à une interrogation de type "recherche par mots-clés" grâce à des fonctions ET, OU, SAUF, NI combinées (adapté d'un programme de l'Institut de Phonétique de Grenoble). Sont en outre activés automatiquement un "mouchard" imprimant la trace des questions posées et des réponses obtenues ainsi qu'une analyse statistique des résultats (nombre d'occurences détectées, durées moyennes, fréquences moyennes, etc.). Par ailleurs, ces résultats peuvent apparaître sur écran ou imprimante ainsi que dans un sous-fichier dans le cas d'interrogations intermédiaires.

3) Exemple d'interrogation

A titre d'exemple, l'interrogation du segment /y/ avec les caractéristiques suivantes :
- appartient à un mot LEXICAL,
- appartient à un mot POLYSYLLABIQUE,
- n'est ni INITIAL, ni FINAL du mot,
- appartient à la DERNIERE SYLLABE du mot,
- se trouve dans un environnement GAUCHE CONSONANTIQUE et dans un environement DROIT VOCALIQUE,
- appartient à un mot SUIVI d'une PAUSE NON FINALE,

permet d'obtenir la durée de chacun des segments se trouvant dans une position telle que dans :
"les jambes du monsieur # sont blanches..."

La figure 2 présente la sortie imprimée de tels résultats.

Une sous-interrogation supplémentaire du type (cs(y)vo) signifiant que (y) est PRECEDE d'une CONSTRICTIVE SOURDE et SUIVI d'une VOYELLE ORALE pourra fournir une précision accrue de la durée (par item ou moyenne) de ce segment dans cette position.

Figure 1

Structure du codage adopté pour la base de données sur l'exemple de la phrase "Le temps est froid et humide"

Col 1 : No de la ligne courante dans la base.
Col 2 : Durée du phone courant (en ms).
Col 3 : phones précédent, COURANT, suivant (2 caractères ; * = V.O.T. des plosives sourdes).
Col 4 : nature de la syllabe courante (ouverte ou fermée).
Col 5 : type du mot courant (Lexical ou Gramtical).
Col 6 : X-Y ; syllabe courante = Xème sur Y dans le mot.
Col 7 : X/Y ; phone courant = Xème sur Y dans la syllabe.
Col 8 : type des phones précédent, COURANT, suivant.
Col 9 : codage "catégoriel" de la position du mot courant.
Col 10 : Fo début, "milieu", fin de la voyelle courante.
Col 11 : contour mélodique de la voyelle courante (=,-,+).
Col 12 : durée de la syllabe courante (en ms).
Col 13 : codage phonétique de la syllabe courante.

1	2	3	4	5	6	7	8	9	10			11	12	13
107	98.	. (L)EU	a	MG	1-1	1/2	(lv)vo	02					139.	[LEU]
108	41.	L (EU)T	a	MG	1-1	2/2	lv(vo)ps	02	89.	87.	81.	=-		
109	162.	EU(T)*	a	ML	1-1	1/3	vo(ps)ms	10					419.	[T*AN]
110	29.	T (*)AN	a	ML	1-1	2/3	ps(ms)vn	10						
111	228.	* (AN)EI	a	ML	1-1	3/3	ms(vn)vo	10	109.	120.	89.	+-		
112	141.	AN(EI)F	a	MG	1-1	1/1	vn(vo)cs	05	90.	77.	86.	-+	141.	[EI]
113	141.	EI(F)R	a	ML	1-1	1/4	vo(cs)lv	27					384.	[FRWA]
114	66.	F (R)W	a	ML	1-1	2/4	cs(lv)dv	27						
115	111.	R (W)A	a	ML	1-1	3/4	lv(dv)vo	27						
116	66.	W (A)EI	a	ML	1-1	4/4	dv(vo)vo	27	102.	104.	103.	==		
117	119.	A (EI)U	a	MG	1-1	1/1	vo(vo)vo	01	103.	95.	89.	--	119.	[EI]
118	120.	EI(U)M	a	ML	1-2	1/1	vo(vo)nv	18	88.	89.	90.	==	120.	[U]
119	132.	U (M)I	f	ML	2-2	1/4	vo(nv)vo	18					560.	[MIDE]
120	153.	M (I)D	f	ML	2-2	2/4	nv(vo)pv	18	83.	71.	70.	-=		
121	119.	I (D)E	f	ML	2-2	3/4	vo(pv)ev	18						
122	156.	D (E).	f	ML	2-2	4/4	pv(ev)	18	68.	64.	66.	-=		

89	56	N (Y)EI	a	ML	1-2	2/2	nv(vo)vo	16					378	[NYEI]
871	73	L (Y)OE	f	ML	1-2	2/4	lv(vo)vo	22					443	[LYOE]
1585	76	D (Y)O	f	ML	1-2	2/5	nv(vo)vo	46					444	[DYO*]
2393	100	N (Y)EI	a	ML	1-3	2/3	nv(vo)vo	07					339	[NYEI]
2508	96	L (Y)EI	a	ML	1-2	2/3	lv(vo)vo	16					379	[LYEI]
3814	89	S (Y)ON	a	ML	1-4	2/3	cs(vo)vo	08					367	[SYON]
3744	62	S (Y)EU	a	ML	1-2	2/3	cs(vo)vo	14					491	[SYEU]
3962	44	S (Y)EU	a	ML	1-2	2/3	cs(vo)vo	14					368	[SYEU]
5475	90	V (Y)ON	a	ML	1-3	2/3	cs(vo)vo	30					354	[VYON]
5740	61	T (Y)AI	f	ML	2-2	2/4	pv(vo)vo	16					477	[TYAIR]

Figure 2

Réponses obtenues à l'interrogation portant sur la semi-voyelle /y/. Voir texte.

III BASE DE CONNAISSANCES

De l'analyse systématique des données engrangées dans la base décrite ci-dessus, il est facile de constituer un tableau (multidimensionnel) de résultats. Le mot "statistique" est intentionellement omis pour cause d'aveuglement de la discipline stricto sensu : l'interrogation de la base de données est toujours suivie (et, bien entendu, précédée !) d'une critique des résultats fournis. Celle-ci inclut la suppression de valeurs jugées "déviantes", voire "aberrantes", mais également la répartition des moyennes globales en un certain nombre de "sous-moyennes" (à partir d'histogrammes) jugées réalistes et propres à générer une variabilité "naturelle" dans un modèle de rythme ou de mélodie.

Cet ensemble de résultats corrigés et interprétés seront stockés dans une BASE DE CONNAISSANCES prenant la forme d'une matrice définie par des dimensions syntaxiques, lexicales, morphologiques et phonétiques.

Sa mise à jour sera d'abord fonction des "trous" éventuels observés, par analyse de corpus complémentaires. Son utilisation pour un nouveau modèle prosodique ne sera plus, dès lors, que banale affaire d'algorithme...

XXIIèmes Journées d'Etudes sur la Parole – Martigny – 15-19 Juin 1998

Hommage à Christian Benoît – JEP 1987

La structure retenue pour la constitution de la base de connaissances est proposée figure 3 sur l'exemple des connaissances apprises en matière de durée des consonnes. L'arborescence figurée suggère un ordonnancement des niveaux de catégorisation "top-down" du syntaxique au phonétique, mais un autre ordre est retenu pour la consultation de la base de connaissances, orienté vers l'interrogation, de manière à ce qu'une requête "pertinente" * restée sans réponse en trouve une, plus générale, au niveau immédiatement supérieur.

Base de connaissances relative aux durées :
Catégorisation des consonnes par arborescence.

	Niveau SYNTAXIQUE	
1	mot précédent	MOT / PAUSE
2	mot suivant	MOT / PAUSE FINALE / PAUSE NON FINALE
	Niveau LEXICAL	
3	mot courant	LEXICAL / GRAMMATICAL
	Niveau SYLLABIQUE	
4	mot courant	MONO- / POLY-SYLLABIQUE
5	position du phonème	INITIALE DE MOT / FINALE DE MOT DANS LA SYLLABE FINALE / AUTRE
	Niveau PHONETIQUE	
6	phonème	VOISE / SOURD
7	phonème	PLOSIVE / CONSTRICTIVE / NASALE LIQUIDE / SEMI-VOYELLE
8	phonème	P/T/K/B/D/G/F/S/CH V/Z/J/M/N/L/R/W/Y
	Niveau CONTEXTUEL	
9	phonème précédent	VOYELLE / CONSONNE
10	phonème suivant	VOYELLE / CONSONNE / E MUET

Figure 3

Les niveaux 6 et 7 se déduisent automatiquement du niveau 8. Des niveaux de connaissances supplémentaires 11 et 12 relatifs au phonème qui suit et/ou précède le phonème courant sont évidemment toujours accessibles (cf figure 1).

En effet, les interrogations sont hiérarchisées ;
- si, par exemple, la durée d'une consonne est demandée ainsi :
1. la consonne appartient à un mot lexical,
2. ce mot est monosyllabique,
3. la consonne est initiale du mot,
4. ce mot est précédé d'un autre mot,
5. ce mot est suivi d'un autre mot,
6. la consonne est dans un environnement gauche et droit consonantique,
- et si aucune occurence ne répondait à ces six conditions, alors la durée retenue pour la consonne serait celle obtenue au terme de la question 5 (voire de la 4).

* Théoriquement, la catégorisation présentée figure 3 envisage 9072 classes possibles pour les durées de consonnes. Après suppression des impossibilités (ex : mot grammatical polysyllabique terminé par /w/), il reste encore bon nombre de possibilités correspondant à des questions "stupides", ce qui laisse suffisamment de redondance pour la répartition de nos 2851 consonnes par locuteur.

IV CONCLUSION

Le but général de cette étude concerne bien évidemment l'établissement de règles prosodiques pour la synthèse.

A court terme, l'objectif est d'extraire du corpus analysé et de valider à la synthèse une valeur de durée pour chaque segment (pause comprise) dans toutes les positions contextuelles possibles en français, ainsi que des contours mélodiques pour toutes les voyelles des mots (pour le français, voir en particulier les travaux antérieurs de [3], [4], [5]).

A plus long terme, l'idée est d'inclure la base de connaissances apprises d'un locuteur dans le système même de synthèse de parole : pour une phrase à synthétiser, il faudra rechercher d'abord une durée moyenne qui corresponde à la situation phonétique, lexicale et syntaxique du segment (ou de la syllabe), ensuite les mélodies de mot qui répondent à la longueur et à la position gauche et droite du mot à synthétiser trouvées dans le corpus grâce au codage "catégoriel" du mot courant et du mot précédent.

Comme plusieurs contours de mot répondront souvent à la demande, il sera alors possible de sélectionner aleatoirement un contour parmi ses pairs, de l'injecter à la synthèse une fois dépouillé de ses évolutions fréquentielles, mais muni de ses variations relatives à la fois dans le mot et en frontière de mot. Seule restera, en fréquence absolue, l'information mélodique représentative du mot initial de phrase, lui aussi extrait du corpus parmi plusieurs possibilités.

Il est légitime d'espérer une amélioration du naturel grâce au respect des contours de mot et des écarts fréquentiels entre mots, comme par la variabilité mélodique introduite aléatoirement. Sans doute, cette amélioration attendue permettra-t-elle de contrebalancer le coût de la recherche.

C'est dans cet esprit, en tous cas, qu'est envisagée la poursuite de cette étude. C'est également dans le sens d'une extension du corpus à d'autres structures de phrase, à d'autres énoncés, à d'autres locuteurs, à d'autres langues.

V BIBLIOGRAPHIE

[1] EMERARD F. et LIENARD J.S. (1980), "Quelques aspects de la traduction phonétique et prosodique du texte écrit", Revue d'Acoustique, 53, 97-101.
[2] MONNE J. (1983), "Programme de calcul de spectrogramme numérique", in "Traitement du Signal de Parole", ENST, Paris.
[3] BAILLY G. (1983), "Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération", Thèse de Docteur-Ingénieur, INPG, Grenoble.
[4] O'SHAUGHNESSY D. (1984), "Design of a Real-Time French Text-to-Speech System", Speech Com., 3, 233-243.
[5] SORIN C. et BARTKOVA K. (1984), "Synthèse de plusieurs styles d'élocution : invariance et variantes prosodiques", 13èmes J.E.P., Bruxelles.

REMERCIEMENTS

A Louis-Jean BOE, pour nous avoir transmis ses programmes d'interrogation de base de données et d'acquisition sur tablette, et pour avoir intégré à sa programmothèque les modifications que nous y avons apporté en retour...

xvi
XXIIèmes Journées d'Etudes sur la Parole – Martigny – 15-19 Juin 1998
Hommage à Christian Benoît – JEP 1987

COMPARAISON DE QUALITE SUBJECTIVE DE TROIS SYNTHETISEURS DE PAROLE

C. BENOIT, M. BOYER, F. EMERARD, C. HAMON

Centre National d'Etude des Télécommunications
B.P. 40
22301 LANNION Cedex

The overall quality of three LPC synthesizers have been compared by means of a pair comparison test over 16 listeners.
A software simulation, a real-time synthesizer using a TMS 320 board on a PC XT, and a commercialized chip from Texas Instruments were tested at different sampling frequencies on a phonetically balanced list of 10 sentences. The test was performed on LPC coded female speech as well as on male synthetic speech obtained by diphone concatenation and a CNET prosodic model.
The same LPC frames files were used as input for the three synthesizers.
The results of the test are presented here. Furthermore, their analysis suggests that the results obtained from a pair comparison using double presentation can be predicted from the results obtained with a single presentation.

I INTRODUCTION

Le Département RCP du CNET a lancé depuis début 87 une étude sur l'"évaluation des systèmes à diffusion de parole sur le réseau téléphonique" dont les buts sont, en interne, la comparaison de qualité de différents synthétiseurs, l'établissement des conditions de recette de nouveaux systèmes, et, en collaboration avec les partenaires européens du projet ESPRIT "SAM", de proposer et tester des méthodes d'évaluation "standard" facilement transportables (voir l'article de DOLMAZON et al. dans ces mêmes J.E.P. [1]).

Préalablement à cette étude, le CNET disposait d'un système de synthèse de parole à partir du texte par concaténation de diphones (SORIN et al. [2]). La partie filtrage numérique par prédiction linéaire existe en laboratoire depuis plusieurs années sous forme de simulation sur un PDP 11/34 muni d'un processeur vectoriel AP120 (STELLA, [3]). Depuis 1986, une version commercialisée de ce synthétiseur est implantée sur PC muni de carte TMS320.

Il était donc important pour les besoins internes du CNET de comparer la qualité globale de ces deux systèmes (*).

En outre, afin de les situer par rapport à un système "du marché", la comparaison de qualité a été étendue au circuit 50C40 de Texas Instrument, incorporant son logiciel microprogrammé.

* Nous entendons ici par "système" l'ensemble matériel, logiciel et les paramètres de timbre constituant la chaîne transformant les trames LPC en signal synthétique.

II LE MATERIEL DE TEST

Notons ici que le synthétiseur 50C40 n'est pas un circuit particulièrement destiné à la synthèse à partir du texte ; les trames filtrées ne pouvant présenter qu'un nombre limité de valeurs de durée.

Il nous a néanmoins paru intéressant de comparer "caeteris paribus" les qualités intrinsèques de ces trois systèmes dans deux conditions spécifiques d'utilisation : l'analyse-synthèse et la synthèse à partir du texte. C'est pourquoi l'ensemble du test, un corpus constitué d'une liste de 10 phrases phonétiquement équilibrée (COMBESCURE, [4]), a été dupliqué : d'une part un jeu de stimuli obtenu par codage LPC de ces phrases lues par une locutrice, d'autre part un jeu symétrique de stimuli obtenu à partir des mêmes phrases synthétisées par concaténation de diphones et application du modèle de prosodie "publicitaire" (SORIN et al., [5]).

D'autre part, les deux systèmes du CNET pouvant fonctionner à trois fréquences d'échantillonnage en sortie : 8, 10 et 16 kHz, et le synthétiseur 50C40 à 8 et 10 kHz, il a été choisi de les comparer deux à deux dans toutes les conditions possibles. Ainsi, ont été préparés 30 fichiers de trames LPC pour l'analyse-synthèse (A/S) et 30 fichiers pour la synthèse à partir du texte (SàpT). Ces mêmes fichiers de trames ont été utilisés pour l'entrée de chacun des trois synthétiseurs.

Les stimuli synthétiques (80 fichiers signaux pour chacune des deux conditions) ont ensuite été rééchantillonnés-filtrés sur le même calculateur (16 bits) puis homogénéisés en intensité.

Afin de comparer les systèmes deux à deux sur chacune des dix phrases dans toutes les conditions permises, nous avons retenu le principe de la double présentation par paires : une paire A-B et une paire B-A, l'ensemble des paires étant présenté aux auditeurs en ordre aléatoire. Il a donc fallu constituer 140 paires pour chacun des deux tests, A/S et SàpT.

Conditions du test en A/S et en SàpT

	TMS	50C40
PDP	8, 10, 16 kHz	8, 10 kHz
50C40	8,10 kHz	

III DEROULEMENT DU TEST

Le test s'est déroulé dans le studio de test du CNET spécialement aménagé, pouvant accueuillir 8 auditeurs simultanément.

Deux groupes de 8 auditeurs naïfs rémunérés ont subi le test en deux séances (A/S et SàpT) séparées d'une journée et découpées en quatre sous-séances chacune (environ 10 minutes chaque sous-séance) :

	matin		après-midi	
1er jour	SàpT	G1	A/S	G2
2ème jour	A/S	G1	SàpT	G2

Dans chaque paire, les deux stimuli étaient séparés de 500 ms. Les paires étant séparées de 5 s entre elles, afin de permettre aux auditeurs de se prononcer sur le choix A ou B, la décision "équivalent" étant autorisée mais vivement déconseillée.

Une feuille présentant le test était remise aux auditeurs avant chaque séance, celui-ci ne débutant qu'au terme d'un pré-test de mise en condition portant sur 5 paires extraites du test.

Les réponses fournies par les auditeurs ont été stockées sur le SOLAR relié aux claviers de réponse du studio ; des programmes de remise en ordre des réponses et de premiers résultats statistiques étant également implantés sur cette machine.

IV ANALYSE DES RESULTATS

Une analyse de la variance a montré l'absence d'effet "auditeurs" et d'effet "phrases" lors de ce test.

Les résultats quantitatifs manifestant les pourcentages de préférence accordés à cahque système sont présentés graphiquement figure 1.

A 8 et 10 kHz, une représentation en "trigramme" a été retenue pour montrer la préférence d'un système sur les deux autres marquée par le plus grand éloignement du système (extrémité du triangle) par rapport à la valeur moyenne de cette préférence. L'écart entre [préfère A] et [100 % - préfère B] correspond à la distance séparant les deux barres perpendiculaires à la droite joignant deux systèmes ; cet écart est très faible, conformément aux consignes données aux auditeurs sur l'usage de la touche "équivalent".

A 16 kHz, condition de comparaison ne concernant que la simulation sur PDP et son adaptation sur PC + TMS320, cette représentation est unidimensionnelle.

Ces résultats globaux montrent une préférence générale marquée pour la simulation (programme du CNET) sur son adaptation temps réel ; la qualité de ces deux synthétiseurs se situant nettement au-dessus de celle du 50C40, tant en SàpT (pour laquelle ce dernier n'est pas adapté) qu'en stockage-restitution.

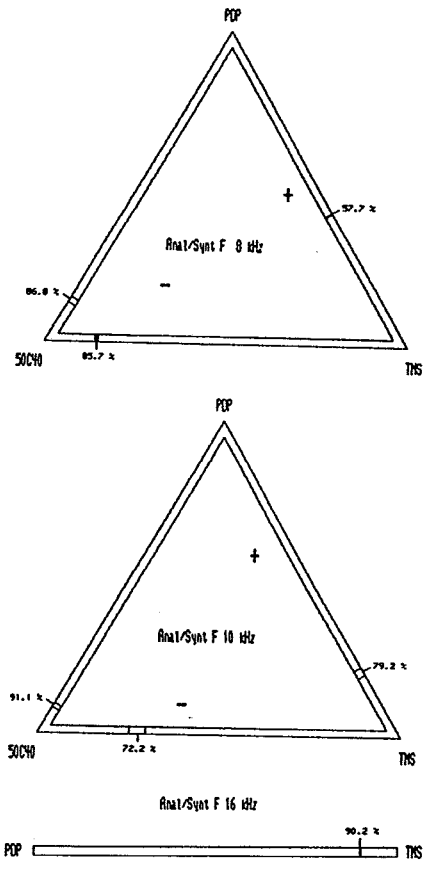


Figure 1-a

Projection des pourcentages de préférence des trois systèmes comparés en analyse-synthèse à 8, 10 et 16 kHz. Ex : à 10 kHz, la qualité globale du "PDP" est préférée dans 79.2 % des cas à celle du "TMS".
Les signes + et - corespondent aux (pseudo) barycentres de préférence et de rejet respectivement ; voir légende Fig. 2.

On observe cependant que l'écart de qualité (ou la certitude du jugement des auditeurs) augmente avec la qualité du codage, de 8 à 16 kHz, dans les deux conditions de test. Ce qui vérifie que les auditeurs sont plus sûrs de leur décision et donc plus sensibles aux défauts intrinsèques d'un système quand celui-ci est moins "extrinsèquement dégradé". Toutefois cette attitude est globale car il est apparu que quelques auditeurs avaient davantage tendance à préférer la qualité "TMS" à la qualité PDP dans les bas débits.

Cette attitude s'est vue confortée par les réponses au questionnaire (informel) remis en fin de dernière séance sur le type de démarche - analytique ou globale - qui a pu guider le choix des auditeurs, selon eux. D'aucuns se sont révélés attirés par un timbre plus aigü, ce qui caractérisait sensiblement le "TMS" par rapport au "PDP". Or, cette préférence tendait à disparaître à 16 kHz car cette "qualité" était alors nettement contrariée par la présence de bruits "métalliques" sur les consonnes sourdes, caractéristique de la tendance "générale" au rejet du synthétiseur "TMS" au profit du "PDP" à cette fréquence.

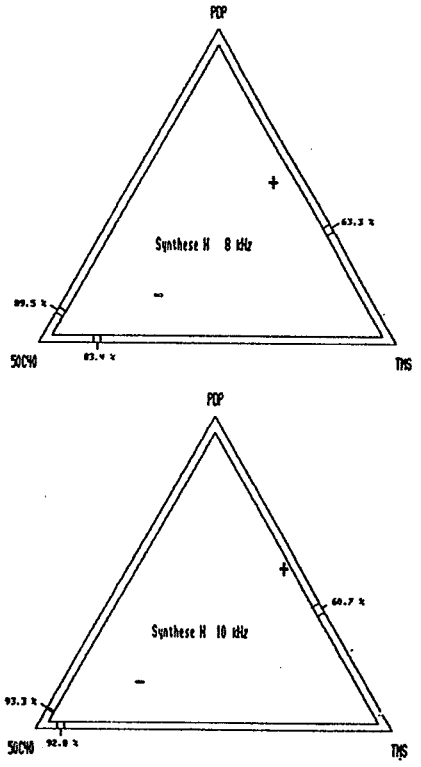


Figure 1-b

Projection des pourcentages de préférence des trois systèmes comparés en synthèse à partir du texte à 8, 10 et 16 kHz. Ex : à 8 kHz, la qualité globale du "TMS" est préférée dans 83.4 % des cas à celle du 50C40.

Il est important de préciser ici que ce résultat et sa critique ont porté leurs fruits puisque depuis lors, ce défaut a été corrigé par une meilleure simulation de la source dans la version temps-réel "TMS" du CNET. Un test de comparaison peut en cacher un autre...

La figure 2 ci-dessous présente les projections des "pseudo-barycentres" des 6 conditions de codage et de débit pondérées par le pourcentage de préférence accordé à chacun des 3 systèmes dans l'espace qu'ils définissent ainsi. L'un de ces points est d'autant plus proche d'un système que celui-ci est préféré aux deux autres dans une condition donnée. Le terme "pseudo-barycentre" est utilisé ici car les poids affectés à chaque système ne sont pas indépendants. L'hexagone figuré en pointillé limite l'espace théoriquement décrit par leurs projections.

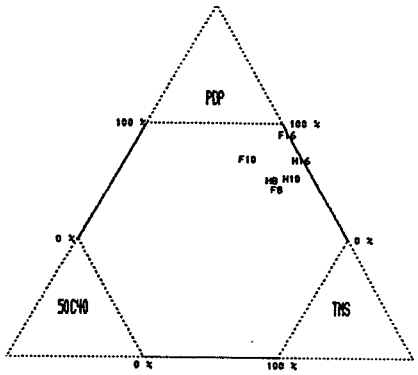


Figure 2

Projections des pseudo-barycentres représentatifs des 6 conditions de test dans l'espace défini par les 3 systèmes testés. H signifie Homme ou SàpT ; F, Femme ou A/S.

V ANALYSE DES REPONSES

Une analyse détaillée du comportement des auditeurs fait apparaître une tendance générale déjà observée par ailleurs : il existe un fort écart entre les réponses suivant que les systèmes A et B sont présentés dans l'ordre A-B ou B-A. C'est en vertu de cette observation classique que l'on est systématiquement amené à doubler le corpus et la durée du test en proposant des doubles présentations : soit A-B-B-A, soit A-B et, séparément, B-A, comme ce fut le cas ici, pour "vérifier" le choix des auditeurs.

Il nous a paru intéressant de comparer les choix des auditeurs suivant qu'un système était présenté en premier ou en second dans la paire de stimuli (stimuli identiques, évidemment, en phrases et en conditions de débit et de codage).

Il apparait ainsi que si globalement, les auditeurs ont tendance à privilégier le système présenté en second, cette réaction est systématique (entendons par là augmenter le score du système présenté en second par rapport au score obtenu quand il est présenté en premier sur les 14 comparaisons binaires possibles) chez 14 sujets sur 16. Elle est à mettre sur le compte de l'effet de mémorisation à court terme. De manière simple, il semblerait que les 5 secondes de temporisation séparant chaque paire permet de remettre la mémoire à l'état "naïf". Le premier stimulus entendu "surprend" par la relative médio-crité de sa qualité, ce qui tend à faire percevoir la présentation suivante, une demi-seconde après, comme moins "choquante", du fait de l'accoutumance opérée à court terme.

La figure 3 présente le diagramme obtenu lorsqu'on observe la différence (en pourcentage) entre le choix 1 et le choix 2 en fonction du pourcentage (moyenné sur les 2 présentations, les 10 phrases et les 16 auditeurs), obtenu par le système préféré dans chacune des 14 comparaisons binaires proposées.

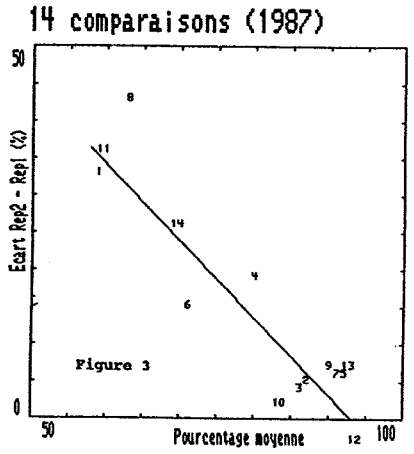


Figure 3

Il apparait nettement une forte corrélation (r=-0.92) entre ces deux observations. Ce résultat n'est pas absolu, puisque la variation est limitée à un quart de plan, les deux variables n'étant pas totalement indépendantes. Néanmoins, il est clair qu'une tendance (sinon une loi) se dessine, permettant de prédire l'écart entre les réponses A-B et B-A en fonction du résultat global. Autrement dit, il est possible de prédire le pourcentage obtenu en double présentation (N %) à partir des réponses à la seule présentation A-B (N1 % obtenu par le système A) suivant la régression linéaire :

$$N = a\,N1 + b$$

N B N1 est supposé > 50% sinon raisonner sur 100%-N1. La tendance observée ici est :

$$N = 2/3\,N1 + 32$$

Il restait à confronter l'observation de cette tendance à des résultats obtenus lors d'autres comparaisons par paires "double présentation".

Un test similaire a été mené en 1984 par F. EMERARD (non publié) sur un corpus identique, afin de comparer les qualités globales de trois synthétiseurs à partir du texte. L'analyse des réponses fournies permet de constater une tendance identique (voir figure 4), même si elle paraît légèrement moins nette. Toujours une préférence globale, caeteris paribus, pour le système présenté en second, ainsi qu'une forte corrélation (r=-0.87) entre l'écart des réponses (A-B, B-A) et leur moyenne.

La tendance à une prédictibilité linéaire est encore ici vérifiée avec, toutefois, des coefficients différents :

$$N = 6/7\,N1 + 12$$

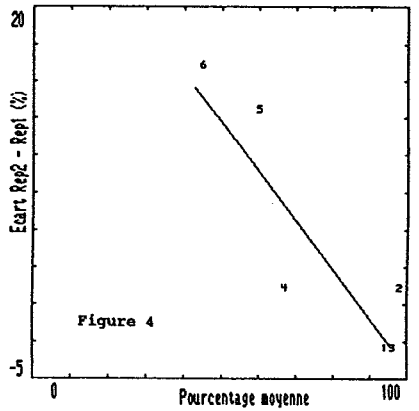


Figure 4

VI CONCLUSION

La validité de tels résultats doit être confortée par d'autres tests. Si elle se trouvait vérifiée, une économie substantielle dans le coût des tests (nombre de stimuli à préparer, durée de la procédure...) pourrait être ainsi réalisée. Or, ce critère est destiné à prendre une place prépondérante dans un avenir très proche.

VII BIBLIOGRAPHIE

[1] J.M. DOLMAZON, C. BENOIT, J.L. GAUVAIN et G. PERENNOU (1987), "Le projet européen "SAM" : évaluation multilingue et dispositifs d'entrées/sorties vocales". 16èmes J.E.P., Hammamet

[2] C. SORIN, R. DESCOUT, C. BENOIT, F. EMERARD, C. FLUHR, D. LARREUR, J.L. LE SAINT MILON, E. MOULINES et R. PERON (1987), "Text-to-Speech Synthesis in the French Electronic Mail Environment", Conference on Speech Technology, Edinburgh.

[3] M. STELLA (1985), "Speech Synthesis", in "Computer Speech Processing", ed. par F. FALLSIDE et W.A. WOODS, Prentice Hall, 421-460.

[4] P. COMBESCURE (1981), "20 listes de dix phrases phonétiquement équilibrées", Revue d'Acoustique, 56, 34-38.

[5] C. SORIN, D. LARREUR et R. LLORCA (1987), "A rhythm-based Prosodic Parser for Text-to-Speech Systems in French", Proceed. of ICPS, Tallin.

XVIIIèmes Journées d'Études sur la Parole
Montréal (Québec), Canada, 28-31 mai 1990

Mesure subjective de la redondance contextuelle : un indice pour quantifier la complexité linguistique

Christian BENOIT

Institut de la Communication Parlée
U.R.A. CNRS N°368 - INPG/ENSERG - Université Stendhal
BP 25 X - 38040 GRENOBLE Cedex FRANCE

RESUME

Un test d'intelligibilité a permis d'évaluer différents synthétiseurs français à partir du texte à l'aide de phrases "sémantiquement imprédictibles". Les distributions de réponses des auditeurs montrent une forte relation entre le pourcentage des phrases et celui des mots correctement retranscrits. Le rapport de leurs logarithmes apparaît comme un indice robuste de la complexité d'un message oral. Des données extraites de la littérature confirment l'hypothèse selon laquelle plus une phrase contient d'informations contextuelles (sémantiques, syntaxiques, etc.), plus faible est cet indice. Il pourrait être relié au nombre d'unités de décision traitées par un auditeur à l'écoute d'une phrase. Les synthétiseurs distordent ici la compréhension des phrases. Or, les erreurs ne suivent pas la loi binomiale qui régirait leur distribution si l'on considérait un modèle simplifié où les unités testées (mots ou syllabes) auraient la même probabilité d'être correctement identifiées. L'analyse des divergences entre l'observation et la théorie issue de ce modèle simpliste montre clairement que la redondance linguistique corrige des mots "*a priori* incompris" et entache d'erreur des mots "*a priori* compris". Ce phénomène de correction/distorsion dans les phrases dépend surtout de leur complexité linguistique, quantifiable par l'indice suggéré, lequel présente également des variations du second ordre dues à la compétence des sujets, à leur entraînement, ou au niveau de dégradation acoustique du message.

1. INTRODUCTION

Si les psycholinguistes ont montré depuis longtemps que l'intelligibilité d'un message oral dépend de la quantité d'information linguistique qu'il véhicule [Miller et al., 1951], la redondance contextuelle n'a cependant jamais été quantifiée.

Il est souvent légitime de considérer une phrase "bien formée" (au sens où l'entend Ruwet [1967], p. 20) comme un ensemble hiérarchisé de mots issus de différents lexiques grammaticaux et dont les positions dépendent des règles syntaxiques de la langue utilisée. Cette conception est opérante pour décrire les langues écrites. Elle est même incontournable à ce jour lorqu'il s'agit de synthétiser la parole à partir du texte. Mais la communication parlée fonctionne selon un schéma éloigné d'une telle combinaison linéaire idéale d'unités. C'est la raison pour laquelle Miller [1962] suggéra le terme d'"**unités de décision**" dans la perception de la parole. Même cachée, leur existence se manifeste quotidiennement dans le dialogue oral ; elle résulte de décisions successives prises au niveau phonétique, lexical, syntaxique, sémantique et pragmatique. Ces décisions, hiérarchisées, reflètent des stratégies ascendantes et descendantes dont le nombre dépend naturellement de la compétence linguistique de chaque auditeur, comme de la complexité linguistique du message émis - ou perçu, s'il n'a pas été correctement compris.

Il n'est malheureusement pas possible de préciser les frontières (à supposer qu'elles existent...) de telles unités de décision dans le cerveau - ou, plus précisément, dans la mémoire de travail - d'un auditeur dont les performances ne peuvent être évaluées qu'en terme de temps de réaction ou de pourcentages d'unités correctement identifiées. Malgré cette incapacité à les identifier, nous tentons ici d'estimer le nombre d'unités de décision traitées par des sujets au cours d'une tâche de transcription orthographique de phrases courtes, acoustiquement dégradées, et au contenu linguistique contrôlé. Des phrases sémantiquement imprédictibles ont été générées automatiquement à partir d'un petit nombre de structures syntaxiques simples. Ce corpus a ensuite été synthétisé sous différentes conditions acoustiques.

Les résultats de cette évaluation de synthétiseurs et la méthodologie adoptée ont été publiés ailleurs (Benoît, 1989). Des tests similaires ont été menés simultanément en Angleterre (Hazan & Grice, 1989) et en Allemagne à partir d'un corpus "multilingue" comparable (Grice, 1989) dans le cadre d'un projet ESPRIT (Benoît et al., 1989).

Nous présentons ici un deuxième niveau d'analyse des réponses des auditeurs pour lequel les synthétiseurs ont servi à dégrader le **support acoustique** de l'information verbale. L'étude des distributions d'erreurs ainsi observées dans la transcription de phrases synthétisées nous permet ainsi de suggérer un nouveau moyen d'en mesurer la **complexité linguistique**.

2. DESCRIPTION DE L'EXPERIENCE

2.1 Le corpus

Nous avons retenu cinq structures syntaxiques simples n'impliquant pas plus de huit mots (y compris les articles définis). Vingt phrases ont été générées aléatoirement, conformément à chaque structure, par concaténation des monosyllabiques français les plus fréquents extraits de lexiques correspondant à leur catégorie grammaticale. Ce test est une extension d'expériences psycholinguistiques classiques [Miller, 1962 ; Miller & Isard, 1963] et d'un test d'intelligibilité de synthétiseurs [Nye & Gaitneby, 1974].

Il faut souligner que les cinq structures ne comportent pas le même nombre de mots (voir Annexe) et que la morphologie du français introduit parfois des modifications du nombre initial de syllabes dues à l'élision de l'article défini devant une voyelle initiale ou de l'insertion d'un *schwa* entre des structures multiconsonantiques finales et initiales de mots. Le nombre moyen de syllabes par structure est en fait inférieur au nombre théorique.

2.2 Les stimulis

Deux synthétiseurs à partir du texte ont été employés dans cette expérience. Ils utilisent le même dictionnaire de diphones, chacun avec sa propre technique de codage. Les cent phrases ont été synthétisées par chacun sous deux conditions prosodiques : un schéma "constamment plat" ; et la modélisation prosodique automatique propre à chacun d'eux. En outre, le locuteur dont la voix avait servi à la constitution du dictionnaire de diphones a enregistré les mêmes cent phrases (légèrement dégradées par la suite) pour comparer la parole naturelle aux 400 stimulis synthétiques. Un ordre de type "carré latin" a été respecté pour aparier les cinq types de voix avec les cinq structures syntaxiques sur cinq bandes magnétiques contenant cent stimulis chacune (mélangés pseudo-aléatoirement).

2.3 Le test

Cinq groupes de quatre sujets ont écouté, à travers un casque, dans une chambre sourde, l'une des bandes magnétiques, à chacune des cinq sessions constituant le test. Le même carré latin que ci-dessus a servi à ordonner la présentation des cinq bandes magnétiques aux cinq groupes d'auditeurs au cours des cinq sessions, ce qui permettait de calculer n'importe quelle score moyen sur les synthétiseurs, les syntaxes, les sessions et les groupes d'auditeurs. Les auditeurs, rémunérés, étaient naïfs en parole synthétique et n'étaient pas informés du contenu linguistique du test. Tous étaient de jeunes étudiants français, sans déficience auditive, et compétents en orthographe comme en écriture manuscrite, de façon à homogénéiser les performances linguistiques. Après une courte phase d'adaptation aux types de phrases et de synthétiseurs, il leur a été demandé de retranscrire sur une feuille de réponse ce qu'ils avaient entendu... ou cru entendre.

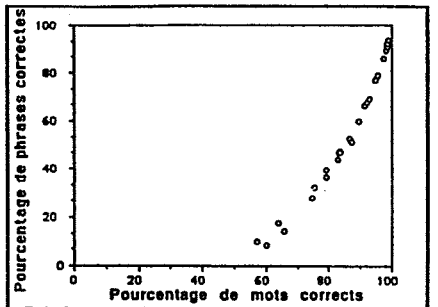
3. RESULTATS

Nous ne fournissons pas ici les résultats détaillés de cette expérience, publiés ailleurs (Benoît, 1989). Rappellons cependant que les cinq conditions acoustiques se répartissent d'une intelligibilité très pauvre à une clarté quasi-totale. Un effet d'apprentisage a été nettement observé, dû en particulier au fait que les cent phrases étaient représentées à chaque session, bien que sous des conditions acoustiques différentes.

3.1. Comparaison des méthodes de calcul des scores

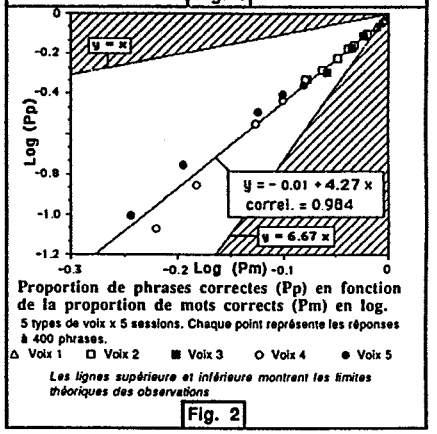
Pour l'ensemble des réponses des 20 auditeurs aux 5 sessions présentant 100 stimulis chacune, ont été calculés les nombres de mots corrects (66700 au total) et de phrases correctes (10000 au total). Tous les mots, articles y compris, ont été considérés. Ils devaient être retranscrits homophoniquement, appartenir à la bonne catégorie grammaticale et être à la bonne position dans la phrase (i.e. position correcte de la syllabe). Fautes, insertions et omissions ont été sanctionnées. Une phrase était correcte si elle ne comportait aucune erreur. Les proportions de mots corrects (Pm) et de phrases correctes (Pp) pour chaque conditon acoustique à chaque session, toutés syntaxes confondues, sont projetées Fig. 1. Une relation exponentielle apparaît clairement entre les deux modes de calcul des scores. Aussi avons-nous représenté Fig. 2 les logarithmes de ces proportions. Il en ressort une forte corrélation et une relation, linéaire cette fois-ci, entre les logarithmes des proportions de mots corrects et ceux des phrases correctes. Le résultat expérimental fournit la relation suivante :

$$\text{Log (Pp)} = r.\text{Log (Pm)} \Leftrightarrow \text{Pp} = \text{Pm}^r \quad (\text{avec } r = 4.27)$$



Relation entre les méthodes de calcul des scores
Chaque point représente 400 phrases ou 2700 mots (20 réponses par 20 auditeurs) sous 25 conditions (5 voix x 5 sessions), toutes structures syntaxiques confondues.

Fig. 1



Proportion de phrases correctes (Pp) en fonction de la proportion de mots corrects (Pm) en log.
5 types de voix x 5 sessions. Chaque point représente les réponses à 400 phrases.
△ Voix 1 □ Voix 2 ■ Voix 3 ○ Voix 4 ● Voix 5
Les lignes supérieure et inférieure montrent les limites théoriques des observations

Fig. 2

3.2 Prédiction des erreurs par la loi binomiale

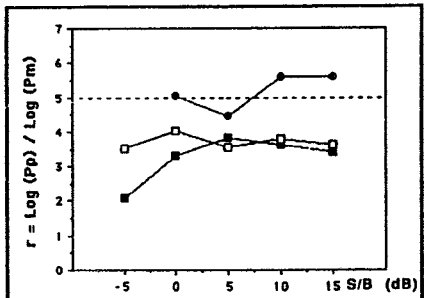
En faisant une approximation grossière, il est possible de considérer un modèle simplifié de phrase dont les unités - c'est-à-dire les mots ou les syllabes - quelles que soient leur position et leur catégorie grammaticale, auraient la même probabilité d'être correctement comprises. Dans un tel cas, une phrase se comporterait comme une suite d'unités dans laquelle une dégradation acoustique (le synthétiseur, ici) introduirait des erreurs de transmission. Si la probabilité qu'une unité soit correctement comprise est pu, la probabilité p(k) qu'une séquence de n unités contienne k unités correctes est alors donnée par la Loi Binomiale :

$$p(k) = pu^k \cdot (1 - pu)^{n-k} \cdot C[n,k]$$

Si k = n, toutes les unités sont correctes. Notons ps = p(n) : probabilité que la séquence soit correcte. Donc, ps = pu^n. Il apparaît ainsi une divergence entre le résultat observé Pp = Pm^r et l'hypothèse simplificatrice ps = pu^n, puisque r = 4.27 < n = 6.67 où n est le nombre moyen de mots par phrase dans notre corpus. Nous devons alors en conclure que le modèle binomial est trop simple quand il suppose que toutes les unités sont indépendantes, et/ou que les mots ne sont pas la bonne unité de perception à mesurer... Du fait des contraintes syntaxiques, et donc de l'information contextuelle, les mots ne sont pas indépendants, ni leurs erreurs de compréhension, tandis que le la loi binomiale suppose des unités aux probabilités indépendantes. Une question se pose alors : la dépendance entre les mots peut-elle être mesurée ? Nous essayons d'y répondre par le postulat suivant : la différence entre n et r peut être considérée comme un indice de la dépendance contextuelle des mots dans la phrase, où r = Log (Pp) / Log (Pm) représente la ***complexité linguistique d'une phrase***, en "équivalents - mots indépendants", ou en unités de décision...

3.3 Comparaison avec des données de la littérature

Miller et Isard (1963) ont évalué l'intelligibilité de phrases grammaticales (i.e. à contenu sémantique), de phrases sémantiquement imprédictibles et de chaînes agrammaticales (i.e. constituées de mots désordonnés). Le même vocabulaire était utilisé dans chaque type de phrases. Les pourcentages de mots corrects et de phrases correctes ont été mesurés sous cinq conditions de dégradation par addition de bruit. Nous avons calculé les rapports de leurs logarithmes qui sont présentés Fig. 3 en fonction du niveau de dégradation. L'indice r observé dans les deux cas de phrases syntaxiquement correctes est systématiquement inférieur à celui obtenu pour les chaînes agrammaticales, lequel est sensiblement égal au nombre initial de mots testés (cinq). De surcroît, la valeur de l'indice pour les phrases grammaticales est inférieur à celui observé pour les phrases sémantiquement imprédictibles dans les pires conditions de dégradation. Ce qui confirme l'idée que cet indice reflète la redondance contextuelle des phrases... sinon le nombre d'unités de décision nécessaire à leur compréhension !

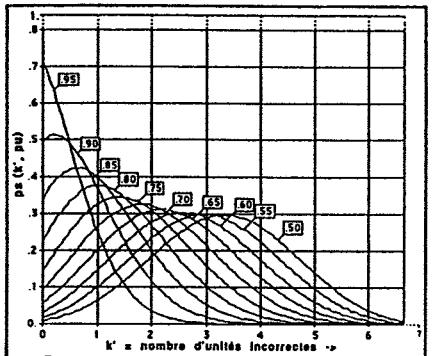


Résultats de Miller & Isard (1963). L'indice r a été recalculé en fonction de la dégradation.

Fig. 3

3.4 Analyse de la distribution des erreurs

Afin d'analyser nos résultats plus en détail, nous avons dû tout d'abord étendre la loi binomiale au cas où le nombre d'unités n'est pas entier, comme c'était le cas dans notre test où nous ne pouvions accéder qu'à un nombre moyen - et réel - de mots par phrase, quelle que soit la structure syntaxique considérée. Pour y parvenir, nous avons extrapolé la factorielle utilisée dans le calcul des combinaisons C[n,k] en la remplaçant par la fonction Gamma qui permet de traiter des valeurs réelles de n et de k. La Fig. 4 présente les courbes continues ainsi obtenues pour différentes valeurs de pu, avec n = 6.67.



Courbes obtenues à partir de la Loi Binomiale étendue à des valeurs réelles du nombre d'unités pour différentes valeurs de pu, probabilité de chaque unité. ps = f(k'), probabilité qu'une séquence de n unités contienne k' erreurs, avec n = 6.67 ici.

Fig. 4

Les mots n'ont pas été testés en présentation isolée. Des valeurs plus réalistes auraient ainsi pu être affectées à Pm, puisque le Pm mesuré ici tient déjà compte *per se* de différentes corrections et distorsions introduites par la redondance contextuelle. Nous n'avons donc pu qu'observer comment les erreurs *a posteriori* sur les mots affectent les erreurs *a posteriori* sur les phrases.

Nous avons calculé les lois binomiales théoriques B(n',pm) approximant au mieux les 25 distributions de réponses des 20 sujets sur 16 phrases sémantiquement imprédictibles, pour chaque structure syntaxique et chaque présentation acoustique. Deux exemples sont présentés Fig. 5. Dans trois conditions acoustiques sur cinq, r et n' ne diffèrent jamais de plus de 0.2 %, quelle que soit la structure syntaxique. Ceci prouve que le rapport des logarithmes correspond rigoureusement au nombre d'unités théoriques qui auraient fourni la distribution d'erreurs la plus proche de celle observée expérimentalement.

4. DISCUSSION

La distribution des erreurs dévie de plus en plus d'une distribution binomiale à mesure que la dégradation augmente. Cela peut signifier que le modèle est irréaliste, mais cela peut aussi être dû au fait qu'il n'est possible d'observer que les distributions de n mots incorrects alors qu'il nous faudrait observer les distributions de n' unités de décision incorrectes...

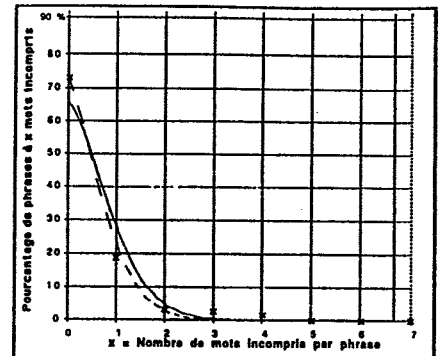


Fig. 5.a : voix synthétique N° 2, syntaxe N° 5

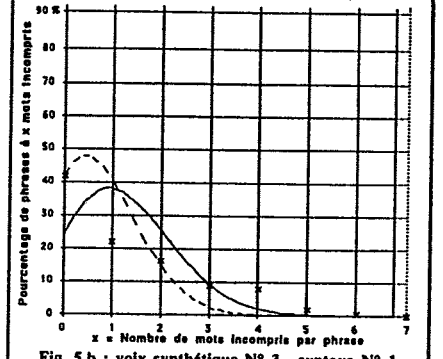


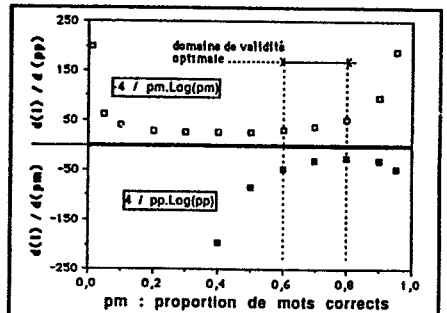
Fig. 5.b : voix synthétique N° 3, syntaxe N° 1

Fig. 5

L'indice r tend vers la valeur de n quand la dégradation diminue. Ceci est naturellement dû au cas limite où la très faible proportion d'erreurs potentielles (ou observées) rend celles-ci strictement indépendantes les unes des autres. Lorsqu'un message est proche de l'intelligibilité totale, la distribution des erreurs résiduelles suit la loi binomiale, et l'observation rejoint la théorie : r est alors égal à n.

Il existe en réalité un niveau optimal de dégradation sous lequel il convient de mesurer l'indice r, comme le suggèrent les résultats, et comme le prouve la logique évidence : au-delà d'une trop forte dégradation, l'information contextuelle se dilue dans la pauvreté du flux acoustique ; au-dessus d'une trop grande clarté, la redondance n'est plus nécessaire à la compréhension d'un message suffisamment simple.

Afin de mieux cerner le domaine de validité de notre indice, nous présentons sur la Fig. 6 l'évolution de sa dérivée par rapport à pm et à pp en fonction de pm pour un nombre donné d'unités de décision théoriques (r = 4 sur la figure). On voit alors nettement qu'en-dessous de 50 % de mots corrects, l'indice r diminue fortement pour une faible augmentation de la proportion de phrases correctes ($\Delta r = 1$ pour une variation de 1% de phrases correctes, à 40 % de mots corrects observés), de même qu'il augmente fortement pour une faible augmentation de plus de 80 % de mots corrects ($\Delta r = 1$ pour un pourcentage de mots corrects observés variant de 90 à 91 %). Le niveau optimal se situe donc sur une plage limitée (entre 60 et 80 % de mots corrects pour un nombre d'unités de décision de l'ordre de 4) au-delà de laquelle la mesure de l'indice r nécessite un nombre élevé de tests subjectifs, en phrases, en répétitions, ou en nombre d'auditeurs pour être précise.



Dérivées de Log (pp) / Log (pm) par rapport à pm et à pp, en fonction de pm

Fig. 6

Dans l'analyse de nos résultats, nous avons comparé nos observations à un modèle dans lequel toutes les unités auraient eu la même probabilité d'être entachées d'erreurs. Nous avons affecté à cette probabilité supposée la valeur moyenne observée dans notre expérience sur un nombre de phrases dépendant de la distribution étudiée (quelques dizaines à quelques centaines). Or, cette valeur correspond à la moyenne arithmétique des erreurs observées. Il est plus réaliste d'imaginer des variations de ces proportions d'erreurs en fonction de la complexité phonotactique de chacun des mots, de leur densité de voisins phonétiquement "proches" dans le lexique, ou encore de leur catégorie grammaticale, voire de leur position dans la phrase. Si la proportion d'erreurs de chacun des constituants de chacune des phrases avait été mesurée, nous aurions alors pu comparer nos observations à des distributions probables d'erreurs calculées à partir des probabilités propres à chaque unité. La valeur particulière correspondant à la probabilité théorique ps qu'une séquence de n unités soit entièrement correcte nous aurait ainsi été fournie par le produit des n probabilités élémentaires : $ps_g = \Pi[i=1,n]\ pu_i$, alors que la loi binomiale utilisée ci-dessus calcule cette probabilité de séquence correcte à partir de la proportion moyenne de mots corrects dans la phrase : $ps_a = \{\Sigma[i=1,n]\ pu_i / n\}^n$. Les racines ènièmes de ces deux probabilités correspondent en fait respectivement aux moyennes **géométrique** et **arithmétique** des probabilités élémentaires. Or, *la moyenne géométrique est toujours inférieure à la moyenne arithmétique.*

D'une part, la différence entre la proportion observée de phrases entièrement correctes (Pp) et la probabilité théorique de leur apparition (ps_g) aurait ainsi été amplifiée. D'autre part, un calcul plus réaliste de notre indice de complexité (également basé sur la moyenne géométrique des proportions observées de mots corrects) aurait fourni une valeur r' (= Log (Pp) / Log (Pm_g), avec $Pm_g = \sqrt[n]{\Pi[i=1,n]\ pu_i}$) encore inférieure à celle que nous avons obtenue ici (r = Log (Pp) / Log (Pm_a), avec $Pm_a = \Sigma[i=1,n]\ pu_i / n$). Si les proportions intrinsèques de chaque mot avaient été utilisées dans cette expérience, un tel calcul aurait donc, non seulement confirmé, mais encore amplifié les différences entre les indices de complexité linguistique mesurés pour les phrases redondantes et pour les chaînes de mots indépendants.

5. CONCLUSION

Ce travail préliminaire suggère que la redondance contextuelle peut être quantifiée. Si sa validité était confirmée, l'indice proposé serait d'un grand intérêt, pour la psycholinguistique comme pour l'évaluation de synthétiseurs de parole, en tant qu'instrument de contrôle de la complexité linguistique. Mais avant cela, de nouvelles expériences doivent être menées afin d'observer l'évolution des pourcentages de mots corrects entre présentations isolées (proportions intrinsèques et *a priori*) et présentations en contexte (proportions *a posteriori*) sous différentes conditions de dégradation. Il est également souhaitable d'évaluer la robustesse de l'indice sur des phrases grammaticales et des chaînes agrammaticales dépassant le nombre "*magique*" de sept ± deux mots, ce qui permettrait de vérifier ou d'infirmer l'hypothèse selon laquelle cet indice serait relié au nombre d'unités de décision prises par les auditeurs. Il n'est en effet pas utopique d'envisager un seuil de complexité au-delà duquel l'absence de redondance linguistique rendrait toute tâche de retranscription correcte impossible, du fait du trop grand nombre d'unités (indépendantes) à traiter par les auditeurs, alors que la présence de la seule syntaxe fournirait suffisamment d'informations contextuelles aux sujets pour que ceux-ci puissent retranscrire correctement certaines phrases et nous permettent ainsi de mesurer le nombre d'unités de décisions auxquelles ils ont eu affaire...

Remerciements

Ce travail a été subventionné par le projet ESPRIT N° 2589 et par le Centre National de la Recherche Scientifique. Nous tenons à remercier les partenaires européens du projet *SAM* pour les nombreux et fructueux échanges scientifiques qui nous ont permis d'avancer dans cette recherche, ainsi que Pascal PERRIER et Jean-Luc SCHWARTZ, pour leurs critiques et commentaires.

Références bibliographiques

BENOIT C. (1989), "Intelligibility test for the assessment of French synthesizers using Semantically Unpredictable Sentences", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.7.1.-1.7.4., Noordwijkerhout, Hollande.

BENOIT C., **ERP** A. van, **HAZAN** V., **GRICE** M. and **JEKOSH** U. (1989), "Multilingual synthesizer assessment using Semantically Unpredictable Sentences", Proceedings of EUROSPEECH'89 Conference, 633-636, Paris.

GRICE M. (1989), "Syntactic structures and lexicon requirements for Semantically Unpredictable Sentences in a number of languages", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.5.1.-1.5.4., Noordwijkerhout, Hollande.

HAZAN V. & **GRICE** M. (1989), "The assessment of synthetic speech intelligibility using Semantically Unpredictable Sentences", Proceedings of ESCA Workshop on "Speech Input/output Assessment and Speech Databases",1.6.1.-1.6.4., Noordwijkerhout, Hollande.

MILLER G.A. (1962), "Decision units in the perception of speech", IEEE Tr. on Information Theory, 8, 81-83.

MILLER G.A., **HEISE** G.A. and **LICHTEN** W. (1951), " The intelligibility of speech as a function of the context of the test materials", J. Exp. Psychol., 41, 329-335.

MILLER G.A. and **ISARD** S. (1963), "Some perceptual consequences of linguistic rules", J. Verbal Learning and Verbal Behavior, 2, 217-228.

NYE P.W. and **GAITENBY** J. (1974), "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences", Haskins Lab. Status Rep. on Speech Research, 37/38, 169-190.

RUWET N. (1967), "Introduction à la grammaire générative", Plon, Paris.

ANNEXE : exemples des cinq structures syntaxiques utilisées dans l'expérience

Stx 1) *"La robe entre vers la science rouge."*
Stx 2) *"Le verre vrai ouvre le coin."*
Stx 3) *"Tourne peu la date et la main."*
Stx 4) *"Quand le texte pose-t-il la fille crue ?"*
Stx 5) *"La chose lance le train qui passe."*

INTELLIGIBILITÉ COMPARÉE DU FRANÇAIS DE FRANCE À GRENOBLE ET À ABIDJAN

C. BENOIT [1], C. CHANARD [2], V. RISSOAN [1] & Z. TCHAGBALE [2]

[1] INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368, INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

[2] INSTITUT DE LINGUISTIQUE APPLIQUÉE, UNIVERSITÉ NATIONALE, 08 BP 887, ABIDJAN 08, République de CÔTE D'IVOIRE

Résumé

Le même test d'intelligibilité, subi à Grenoble par vingt auditeurs français, puis à Abidjan par vingt auditeurs ivoiriens, pour qui le français est une langue seconde, a permis de comparer les performances de deux groupes de sujets de même niveau d'étude mais aux compétences linguistiques différentes pour ce qui concerne la compréhension du français parlé en France, ou modélisé comme tel. Le test consistait, pour chaque sujet, en une transcription manuscrite de 500 phrases syntaxiquement simples, mais sémantiquement imprédictibles, et formées des monosyllabes les plus fréquents du français. Les stimulis avaient été synthétisés suivant deux méthodes de codage et deux modèles prosodiques élaborés pour le français "standard" d'une part, et prononcés par un locuteur français avant d'être spectralement dégradées d'autre part. Les scores d'intelligibilité globaux, en nombre de phrases correctement retranscrites, font apparaître une différence marquée entre les performances du groupe français et celles du groupe ivoirien. En revanche, la complexité linguistique perçue (mesurée comme le nombre d'unités de décision indépendantes nécessaires aux sujets pour la transcription de leur réponse : cf. l'article de Benoît aux précédentes JEPs) est indépendante de l'origine des sujets. Une telle observation est riche de conséquences pour les pédagogues comme pour les Industriels de la Langue Française ayant vocation à travailler avec ou pour la Francophonie en Afrique.

1. INTRODUCTION

Lors des précédentes JEPs (Benoît, 1990b), ont été présentés les résultats d'un test d'intelligibilité de la parole, synthétique et naturelle dégradée, réalisé auprès d'un échantillon de vingt auditeurs, tous étudiants français sans déficience auditive. Une mesure subjective de la complexité linguistique des phrases testées avait également été présentée L'indice proposé (rapport entre le logarithme des proportions de phrases correctes et celui des proportions de mots corrects) avait été utilisé antérieurement par Boothroyd (1988, 1990) pour quantifier une telle complexité, dans des phrases mais aussi dans des mots sans signification, tant auprès de sujets à audition normale qu'auprès de déficients auditifs. Ces travaux montrent que cet indice correspond au nombre d'unités de décision indépendantes traitées par les auditeurs pour tenter de comprendre une structure linguistique, mot ou phrase. En effet, la redondance introduite dans un message, compte tenu des contraintes au niveau du lexique, de la syntaxe, ou de la sémantique est une information mise à profit par les auditeurs pour réduire l'éventail des candidats possibles à la "compréhension" d'une unité linguistique (phonème, syllabe, ou mot, selon les cas et les niveaux de stratégie descendante). Ce phénomène se manifeste tout particulièrement quand la transmission acoustique du message est dégradée : la répartition des erreurs de transcription de messages suffisamment simples montre alors que les auditeurs traitent une structure de N unités linguistiques comme si elle n'était constituée que de P unités indépendantes (avec P < N), lorsque ces N unités sont reliées entre elles par un certain nombre de règles linguistiques (phonotactiques, lexicales, syntaxiques, etc.). La "redondance linguistique" du message peut donc être évaluée à N – P unités indépendantes.

Dans cet article, nous comparons, entre deux communautés francophones géographiquement et culturellement disctinctes, non seulement l'intelligibilité globale de phrases transcrites manuellement, mais aussi la capacité des deux groupes d'auditeurs à tirer profit de cette "redondance linguistique" introduite, dans des phrases sémantiquement imprédictibles, par le respect de règles lexicales et syntaxiques.

Les phrases testées obéissent à cinq structures syntaxiques de base et ont été générées automatiquement par concaténation des monosyllabes français les plus

fréquents dans chacune des catégories grammaticales nécessaires. De telles phrases, appellées "sémantiquement imprédictibles" (Benoît et al., 1989 ; Grice, 1989 ; Benoît, 1990b), limitent l'aide à la compréhension des mots qui les constituent, lorsqu'elles sont présentées aux auditeurs sous forme acoustiquement dégradée, au seul respect de leur existence lexicale et de quelques règles syntaxiques de base en français. La dégradation que nous avons introduite dans ce test était soit due aux performances limitées de différents synthétiseurs de parole dans quatre cas, et à une dégradation artificielle du spectre de parole naturelle dans un cinquième cas. Les cinq conditions de "codage" utilisées permet de couvrir un champ suffisamment vaste de dégradation.

2. LES STIMULIS

Cinq structures syntaxiques ont été retenues, pour leur simplicité et pour leur adaptibilité, à quelques ajustements mineurs près, à l'ensemble des langues indo-européennes (Grice 1989). Des phrases ont ainsi pu être générées automatiquement à partir de lexiques grammaticaux contenant les monosyllabes les plus fréquents dans leur catégorie. Un exemple de "cadavre exquis" ainsi obtenu est présenté au tableau 1 ci-dessous pour chacune des cinq structures:

Tableau 1 : exemples de phrases utilisées dans le test

Structure 1 (forme intransitive)
La robe entre vers la science rouge.

Structure 2 (forme transitive)
Le verre vrai ouvre le coin.

Structure 3 (forme impérative)
Tourne la date ou la main.

Structure 4 (forme interrogative)
Quand le texte pose-t-il la fille crue ?

Structure 5 (forme relative)
La chose lance le train qui pense.

Vingt phrases ont été générées dans chacune des cinq structures. Les 100 phrases ont ensuite été synthétisées sous quatre conditions chacune : un synthétiseur à diphones codés à formants et un synthétiseur à diphones codés PSOLA, tous deux avec leur modélisation prosodique propre, et avec une prosodie constante. Elles ont aussi été lues par le locuteur dont la voix avait servi de référence pour la constitution des deux dictionnaires de diphones. Cinq bandes magnétiques ont alors été enregistrées. Chacune contenait les cinq groupes de vingt phrases d'une même structure syntaxique, présentées sous la même condition acoustique, après mise en ordre aléatoire de leur présentation. Un carré latin a permis d'aparier structures syntaxiques et conditions de codage de façon à ce que les cinq bandes magnétiques de 100 phrases chacune ne contiennent pas de répétition. Une pause de 10 secondes séparait deux phrases consécutives, et une pause d'une minute séparait deux groupes de vingt phrases.

3. LES AUDITEURS

A Grenoble comme à Abidjan, les auditeurs sélectionnés ne présentaient aucune déficience auditive connue. Ils étaient tous étudiants à l'université et avaient une bonne maîtrise de l'orthographe et de l'écriture manuscrite. Ils étaient rémunérés pour participer à l'expérience.

Le français était langue maternelle pour les 20 auditeurs grenoblois ; il était langue seconde pour les 20 auditeurs ivoiriens. Les auditeurs ivoiriens avaient tous poursuivi leur scolarité en français, langue véhiculaire officielle en Côte d'Ivoire.

Notons que, si le français est très répandu et parfaitement maîtrisé par les étudiants ivoiriens, sa spécificité régionale le rend de compréhension difficile pour un auditeur français peu habitué à cette variante. En revanche, les Ivoiriens sont beaucoup plus sensibilisés au français de France, du fait des médias en général, mais aussi de la présence française à l'Université pour les étudiants. Cette évidente dissymétrie de compréhension entre les deux variantes du français n'est pas étudiée ici, puisque nous avons seulement testé l'intelligibilité du français "de France" auprès des deux communautés.

4. LE TEST

Les deux tests se sont déroulés selon un protocole rigoureusement identique : cinq sous-groupes de quatre sujets ont subi cinq sessions au cours de chacune desquelles ils devaient retranscrire manuellement les 100 phrases entendues. Un carré latin avait réparti chacun des cinq groupes dans chacune des cinq sessions de façon à ce qu'une des cinq bandes magnétiques soient entendue à chaque session par un groupe, sans répétition. Les quatre auditeurs, équipés d'un casque de haute qualité, se tournaient le dos. Une brève session de familiarisation préalable leur permettait de prendre connaissance du genre de messages qu'ils avaient à retranscrire, puis du type de dégradation subie, sans autre précision sur le contenu, linguistique ou acoustique, du test. Il leur était recommandé d'être plutôt imaginatif en cas de mauvaise compréhension et de transcrire un trait quand un mot était jugé totalement incompréhensible.

5. LES RÉSULTATS

Les réponses aux 2 (tests) x 20 (auditeurs) x 5 (sessions, ou bandes magnétiques) x 100 phrases — soit 20 000 phrases, ou 130 000 mots environ — ont été saisies dans un fichier ASCII de façon à être analysées automatiquement en fonction de différents

facteurs. Nous renvoyons le lecteur à l'article antérieur de Benoît (1990b) pour une présentation détaillée des résultats chez les auditeurs français. Nous nous limitons ici à une comparaison des deux observations les plus significatives : l'intelligibilité globale des phrases et des mots pour chacun des auditeurs, moyennée sur l'ensemble des conditions acoustiques, syntaxiques et sur les sessions, d'une part ; et l'indice calculé comme le rapport des logarithmes des proportions précédentes, de façon à estimer la complexité linguistique perçue du corpus testé.

5.a Intelligibilité globale : comparaison des performances acoustiques

Nous avons mesuré l'intelligibilité globale moyenne des phrases composant l'ensemble du test, chez chacun des quarante auditeurs, comme le pourcentage d'unités correctement retranscrites. Comme unité, nous avons tout d'abord retenu le mot, celui-ci étant un monosyllabe dans chaque catégorie. Un mot a été considéré comme correct quand il était retranscrit à la bonne place dans la phrase, sous une forme orthographique telle qu'il pouvait appartenir à la catégorie grammaticale correspondante, avec les mêmes attributs morphologiques, et sous une forme homophonique. Ainsi, les mots *mer* ou *mère* étaient considérés comme intervertibles, alors que *mer* et *maire* ne l'étaient pas, ces derniers n'étant pas du même genre, bien qu'étant des noms tous les deux. Il faut noter ici que des mots susceptibles d'appartenir à deux catégories comme *vieux* (nom ou adjectif) n'avaient été retenus que dans le lexique de la catégorie la plus fréquente (adjectif, ici), de façon à désambiguïser l'appartenance d'un mot à la catégorie grammaticale prévue. Enfin, l'autre unité utilisée pour le critère d'intelligibilité était la phrase. Une phrase était considérée comme correcte si elle ne contenait aucun mot incorrect suivant le critère précédent.

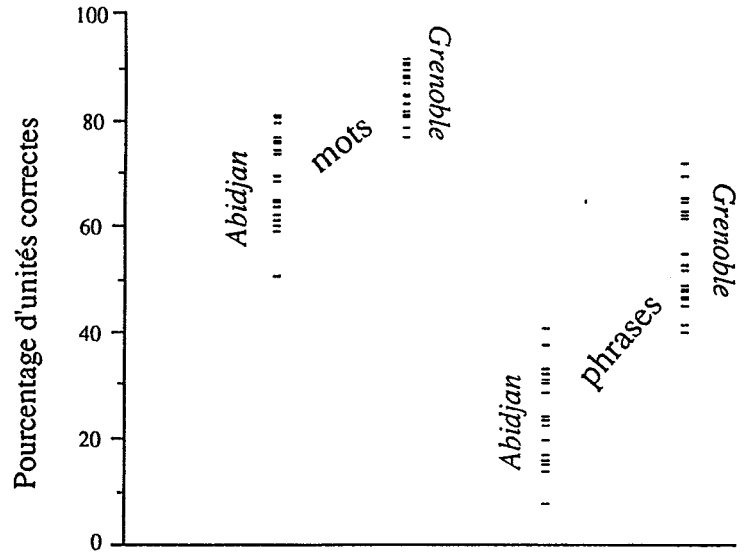


***Figure 1.** Scores moyens d'intelligibilité, en pourcentage de mots corrects (à gauche) et de phrases correctes (à droite) obtenus par les vingt auditeurs ivoiriens (colonnes de gauche) et les vingt auditeurs français (colonnes de droite) sur les 500 phrases retranscrites.*

La Figure 1 présente les scores moyens d'intelligibilité en proportion de mots corrects et en proportion de phrases correctes chez les auditeurs ivoiriens et chez les auditeurs français. La différence de performance globale est très nette entre les deux groupes, puisque seul le "meilleur" des auditeurs ivoiriens obtient un pourcentage de phrases correctes supérieur à celui obtenu par le "pire" des auditeurs français.

Ce résultat global doit toutefois être nuancé en fonction de la dégradation acoustique, voire de la modélisation prosodique. Comme le montre la Figure 2, la parole naturelle, assez faiblement dégradée, permet une discrimination totale entre les deux communautés, alors qu'un léger recouvrement est observé dans le cas du synthétiseur n° 1, et que le synthétiseur n° 2 confond encore plus nettement les deux communautés. Deux raisons peuvent être avancées a priori pour expliquer ce phénomène : les différences régionales d'intelligibilité sont atténuées soit quand l'acoustique est plus dégradée, soit quand la prosodie est moins naturelle. Or, en moyenne sur les 20 auditeurs ivoiriens, le pourcentage d'unités correctes n'est pas significativement modifié d'un modèle prosodique à l'autre (ou à son absence, plus exactement), quel que soit le synthétiseur. A l'inverse, chez les 20 auditeurs français, l'intelligibilité du synthétiseur n° 1 augmente de 58.0 % de phrases correctes, quand il est présenté en "prosodie constante", à 65.4 % quand il est présenté avec sa "prosodie modélisée", tandis que celle du synthétiseur n° 2 ne varie pratiquement pas d'une condition prosodique à l'autre. Comme la parole naturelle n'était présentée aux auditeurs qu'avec sa prosodie naturelle, robuste au type de dégradation acoustique infligé, il ne nous est pas permis de trancher dans cette énigme !

A travers les trois conditions acoustiques testées, il existe une étroite relation entre les qualités du codage acoustique et de la prosodie (du français parlé en France) associée : très bonnes en parole naturelle, médiocres avec le synthétiseur n° 1, et très pauvres avec le synthétiseur n° 2. C'est pourquoi nous nous limiterons ici à conclure, dans l'attente d'expériences complémentaires, soit qu'une forte dégradation acoustique ramène les deux groupes linguistiques à une incompréhension comparable des messages émis ; soit que c'est pour une très grande part l'information prosodique caractéristique d'une communauté de Francophones (ici celle de France) qui fait la différence d'intelligibilité entre deux messages, naturels ou synthétiques.

Il est légitime de supposer que les deux effets existent, et que ceux-ci se sont conjugués dans l'expérience rapportée ici.

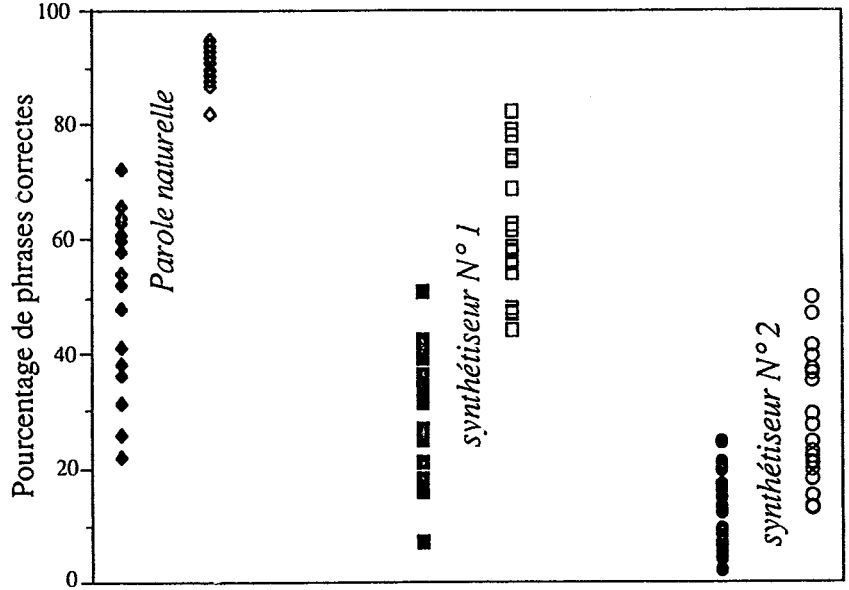


***Figure 2.** Comparaison des pourcentages de phrases correctement retranscrites par les deux communautés selon le type de présentation acoustique: parole naturelle (faiblement) dégradée, et deux types de synthétiseur. Pour ces derniers, les résultats confondent les deux versions prosodie constante et prosodie modélisée.*

Un tel résultat doit évidemment être relativisé : les conditions expérimentales décrites ci-dessus sont (heureusement !) fort éloignées d'une situation de communication usuelle, dans laquelle les messages échangés entre interlocuteurs sont moins imprédictibles, chargés de sémantique et de pragmatique, d'une part, mais aussi moins acoustiquement dégradés ! Ainsi, dans des conditions aussi optimales que celles offertes par la télévision, il est permis de supposer que les messages diffusés "en français de France" aux auditeurs ivoiriens sont aussi bien compris que par les auditeurs français... En revanche, ce résultat établit clairement que les Industriels de la Langue et, avant eux, les chercheurs, doivent encore fournir un gros effort pour que leurs synthétiseurs de parole soient aussi compréhensibles par l'ensemble des utilisateurs francophones potentiels !

5.b Complexité perçue : comparaison des performances linguistiques

Une fois admises les différences intercommunautaires dans leurs performances de "bas niveaux", en matière de décodage acoustico-phonétique, et/ou en matière de décodage prosodique qui mettent en jeu des stratégies essentiellement ascendantes, il était intéressant de comparer leurs performances de plus "haut niveau", en matière de compétence linguistique mettant en jeu des stratégies descendantes.

Afin d'évaluer les différences cognitives des auditeurs, nous avons comparé la complexité linguistique des phrases perçue par chaque communauté. Pour ce faire, nous avons calculé, chez chaque auditeur, le rapport des logarithmes des deux proportions d'unités correctes r = Log (Pp) / Log (Pm).

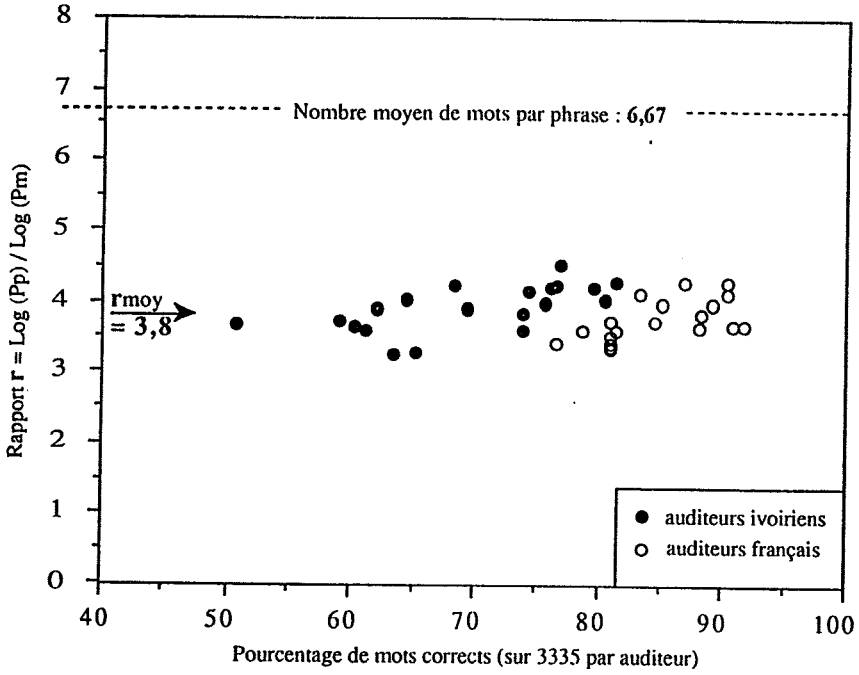


Figure 3. Variation de "l'indice de complexité" r = Log (Pp) / Log (Pm) en fonction de l'intelligibilité globale en pourcentage de mots correctement retranscrits.

Le rapport r est un indice robuste de la déviation entre i) la loi binomiale théorique des distributions d'erreurs sur des mots considérés comme indépendants les uns des autres et ii) la distribution observée des erreurs sur les mots des phrases testées, où des règles syntaxiques simples relient certaines unités entre elles : accord en genre entre le déterminant, l'adjectif et le nom ; accord morphologique entre ceux-ci et le verbe ; ordre d'apparition des mots dans la phrase ; etc.

Boothroyd (1988) et Benoît (1990b) ont montré que le rapport r = Log (Pp) / Log (Pm), dans lequel Pp représente la proportion de phrases correctes et Pm celle des mots corrects sur l'ensemble du test, représente le nombre d'unités de décision prises par un auditeur pour la compréhension des phrases d'un corpus homogène. Cet indice est sensible à de faibles variations des proportions. C'est pourquoi nous l'avons appliqué à l'ensemble des 500 phrases transcrites par chaque auditeur, sans distinction de structure syntaxique, afin d'en augmenter la précision. Il ne reflète donc pas la part prise par chacune des structures dans la "facilitation" à la compréhension des mots pris individuellement, mais l'avantage global apporté par l'ensemble des cinq structures à la compréhension de toutes les phrases. Cette approche holistique n'est pas choquante, dans la mesure où les cinq structures reflètent des syntaxes de base du français, et où les phrases ont sensiblement le même nombre de mots (6,67 en moyenne sur les 500 phrases).

La Figure 3 présente la valeur calculée de cet indice, pour chaque sujet, en fonction de la proportion de phrases correctement retranscrites.

La valeur du rapport des logarithmes reste à peu près constante d'un auditeur à l'autre (3,8 en moyenne contre 6,67 mots par phrase). Tous manifestent une large capacité à utiliser la redondance linguistique des phrases pour les traiter comme si elles n'étaient constituées que de 3,3 à 4,5 "unités indépendantes". Il n'apparaît donc pas de différence significative entre les deux communautés dans leur habileté à utiliser la redondance contextuelle des phrases, quelle que soit leur niveau d'intelligibilité.

7. CONCLUSION

Bien que les conditions dans lesquelles nous avons placé nos auditeurs étaient acoustiquement très dégradées et linguistiquement peu complexes, les observations tirées de l'expérience décrite dans cet article nous permettent d'affirmer que, si les auditeurs ivoiriens ont plus difficulté que les auditeurs français à décoder acoustiquement – ou prosodiquement – des messages dégradés, les deux groupes font preuve d'une même compétence dans leur décodage linguistique.

A notre connaissance, aucune étude comparative n'avait été menée jusqu'à présent sur l'intelligibilité comparée du français de France entre deux communautés francophones. Si de nombreux travaux ont été menés, dans les pays du Nord, sur l'analyse et la synthèse du français parlé au Québec, peu de cas est fait, dans la littérature, des variantes du Français parlé dans les pays du Sud. Or, il n'est pas utopique d'imaginer que l'utilisation de machines parlantes se répendra en Afrique aussi rapidement qu'en France ou au Canada, comme cela s'est déjà produit pour nombre de nouveautés technologiques.

N'ayons donc pas peur de saisir l'occasion des Journées d'Étude sur la Parole pour affirmer qu'il serait linguistiquement teinté d'un certain relent colonialiste francophone de la part des Industriels de la Langue – et plus encore, des chercheurs – si leurs synthétiseurs de parole considéraient davantage les variantes régionales du français comme l'objet d'un futur marché commercial que comme celui d'une étude linguistique. Ils développeraient alors plus un outil de normalisation linguistique qu'un support scientifique à l'enrichissment des connaissances universelles...

Remerciements

Cette étude a été partiellement subventionnée par le projet ESPRIT N° 2589, et par l'Agence de Coopération Culturelle et Technique.

Références bibliographiques

Benoît, C., van Erp, A., Grice, M., Hazan, V., & Jekosh, U. (1989), "Multilingual synthesizer assessment using semantically unpredictable sentences", *Proceedings of the Eurospeech '89 Conference*, Paris, 633-636.

Benoît, C. (1990a), "Mesure subjective de la redondance contextuelle : un indice pour quantifier la complexité linguistique", *Actes des 18e Journées d'Etude sur la Parole*, Montréal, Canada, 159-163.

Benoît, C. (1990b), "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity", *Speech Communication*, 9, 293-304.

Boothroyd, A., & Nittrouer, S. (1988), "Mathematical treatment of context effects in phoneme and word recognition", *Journal of the Acoustical Society of America*, 84, 101-114.

Grice, M. (1989), "Syntactic structures and lexicon requirements for Semantically Unpredictable Sentences in a number of languages", Proceedings of the ESCA workshop on Speech I/O Assessment and Speech Databases, Noordwijkerhout, Hollande, 151-154.

Nittrouer, S., & Boothroyd, A. (1990), "Context effect in phoneme and word recognition by young children and older adults", *Journal of the Acoustical Society of America*, 87, 2705-2715.

SYNTHÈSE À PARTIR DU TEXTE D'UN VISAGE PARLANT FRANÇAIS

P. WOODWARD, T. MOHAMADI, C. BENOIT & G. BAILLY

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

Une analyse antérieure a montré qu'une vingtaine de visages prototypiques (ou visèmes) permet de rendre compte de la diversité des formes maxillo-labiales de la parole non expressive en français. Cet article décrit les choix scientifiques et techniques pour lesquels nous avons opté : l'animation de visage par affichage d'images successives ; le jeu d'images-clé retenu ; l'architecture du système de synthèse audio-visuelle de la parole à partir du texte ; l'implémentation des règles de transcription phonème-visème tenant compte de la coarticulation en français ; et la synchronisation du son et des images synthétiques.

Ce système fonctionne sur un PC muni d'une carte de traitement du signal audio et d'une carte graphique VGA. Les règles décrites ont été élaborées sous le formalisme du compilateur de règles COMPOST, développé auparavant pour la synthèse acoustique à partir du texte.

1. INTRODUCTION

Parmi les théories et les techniques de la perception, la vision s'est affirmée d'emblée comme un domaine de recherche privilégié, et ce pour plusieurs raisons. De tous nos sens, le sens de la vue est celui qui offre la plus grande flexibilité et nous apporte la plus grande quantité d'information sur ce qui nous entoure du fait de sa triple dimension spatiale, énergétique et temporelle. La perception de notre espace, la détection du mouvement des objets qui le composent sont des données humaines essentielles et sont toutes transmises par l'intermédiaire de nos yeux.

Dans le domaine de la communication, l'omniprésence de l'informatique dans notre société fait que nous sommes toujours à la recherche d'une meilleure communication homme-machine. C'est en partie dans cette optique que l'on a commencé à équiper les ordinateurs de voix de synthèse et que l'on manifeste aujourd'hui un intérêt croissant à l'égard de la synthèse de visage parlant. Cette synthèse consiste à doter l'ordinateur d'une animation d'images faciales qui accompagne la voix artificielle.

L'animation de visage est d'autant plus nécessaire que l'intelligibilité actuelle des synthétiseurs de parole est encore très éloignée de celle des humains, et qu'il est bien connu que la perception visuelle de la parole dégradée vient compléter sa perception acoustique par un apport d'information supplémentaire (Sumby et Pollak, 1954 ; Neely, 1956 ; Binnie et al., 1974 ; Erber, 1969, 1975 ; Summerfield, 1979 ; Mohamadi et Benoît, ce volume).

En outre, il est séduisant de viser à améliorer l'impact d'un message en l'*humanisant* au moyen d'un visage, même schématique.

Dans un premier temps, nous avons élaboré une animation de visage basée sur le principe classique de l'affichage d'images successives, celles-ci étant préstockées dans un *dictionnaire de visages*. Cette technique facilite la validation des règles de transcription, bien qu'elle limite les possibilités de visualisation à un jeu fini d'expressions. Notre synthétiseur ne peut donc rendre compte, dans son état actuel, que d'une prononciation *neutre*, inexpressive, qui est néanmoins la condition optimale pour mener des tests de perception de la parole, *et seulement* de la parole...

Le synthétiseur de visage décrit ici vise donc à apporter intelligibilité et agrément à la synthèse de la parole à partir du texte, et à offrir un outil de validation des connaissances en matière de production et de perception de la parole. L'évaluation de son intelligibilité fera l'objet d'une prochaine étude.

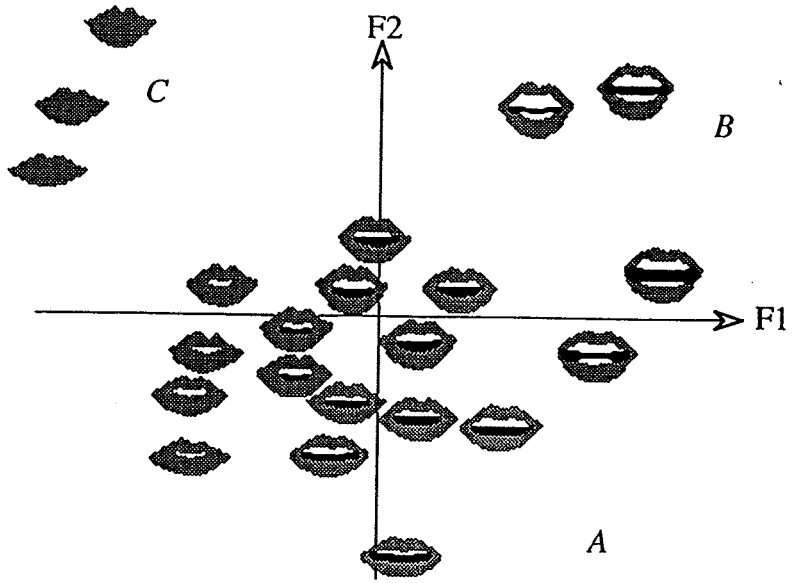


Figure 1.a. Projection factorielle des 22 formes (schématisées) de lèvres et de dents utilisées pour la synthèse audio-visuelle de la parole. Les seules vues de face sont présentées ici (sans le menton) dans un plan essentiellement caractérisé par les jeux de la protrusion du contact labial (C), de l'étirement (A) et de la séparation (B) intérolabiales.

2. LE DICTIONNAIRE DE VISÈMES

Dans un étude récente (Benoît et al., 1992), il a été établi qu'une vingtaine de formes maxillo-labiales prototypiques d'un locuteur permettait de rendre compte de la géométrie des lèvres et du menton, au centre de la réalisation acoustique des 14 voyelles du français en isocontexte, des trois voyelles extrêmes [i, a, y] en différents contextes consonantiques, et des six consonnes [b, v, z, ʒ, ʁ, l] en différents contextes vocaliques. Les 23 classes – ou visèmes – révélées par cette analyse ne prétendent ni à une description fine de l'ensemble des formes maxillo-labiales significativement différentes, ni à une exhaustivité complète de tous les effets contextuels possibles en français. Pour les raisons évidentes, le corpus étudié était limité aux réalisations et aux contextes jouant un rôle majeur sur la dynamique des articulateurs visibles de la parole. Des règles fondamentales sur l'effet général des différents contextes, vocaliques et consonantiques, ont pu être observées à partir des données obtenues (Benoît et al., 1991), ouvrant la voie à leur *extrapolation* à l'ensemble les réalisations possibles. Enfin, les visèmes recensés correspondant aux formes caractéristiques au centre des réalisations acoustiques, une *interpolation* des formes est rendue nécessaire dans des conditions de débit lent où l'affichage d'une image intermédiaire est alors indispensable entre deux visèmes consécutifs : une grammaire des règles d'organisation temporelle de ces formes est également à l'étude en vue d'identifier, parmi les visèmes existants, ceux susceptibles d'être intercalés entre deux visèmes successifs, non contigüs dans l'espace multidimensionnel où ils ont été définis.

La figure 1 présente la projection factorielle (obtenue par analyse des correspondances sur l'ensemble de toutes les réalisations et des 14 paramètres géométriques mesurés sur celles-ci) des 22 visèmes caractéristiques de la parole (la position *repos – lèvres fermées* n'est pas représentée) et des trois paramètres les plus discriminants (A : écartement intérolabial ; B : séparation intérolabiale ; C : protrusion du contact labial). La figure 1.a présente la vue de face des lèvres et des dents correspondant aux réalisations des visèmes représentés par leur symbole à la figure 1.b. La liste par extension de ces mêmes visèmes est donnée dans le tableau 1. On notera, sur ces figures, l'important effet du contexte [ʒ] sur les voyelles [i] et [a], le long du facteur F2 (combinaison de protrusion et de séparation des lèvres), ou celui du contexte [z] sur [a], par exemple, qui *trans-forme* la forme du [a] "tenu" en celle d'un [i] "tenu" (par rapprochement des lèvres).

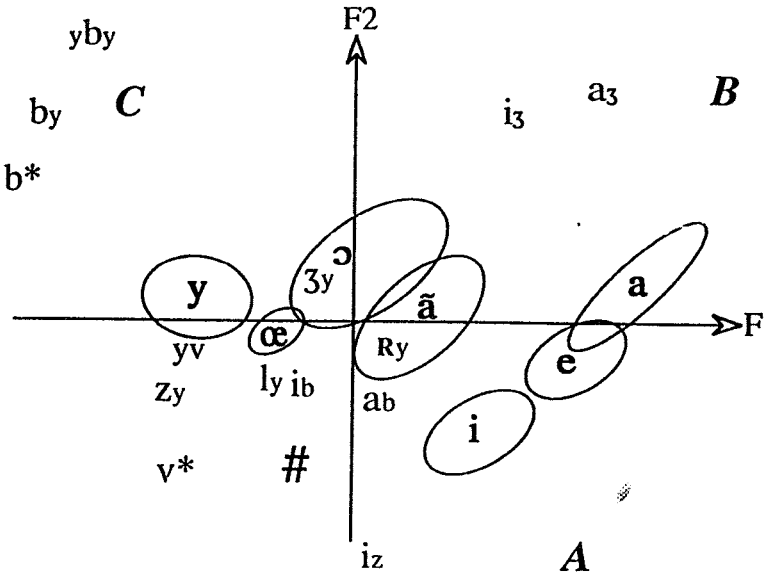


Figure 1.b. Projection factorielle des 22 visèmes définis par leurs symboles (cf. leur définition dans le tableau 1. ci-dessous). Les ellipses de dispersion figurent les nuages de réalisations des visèmes vocaliques à 90 % de confiance.

Tableau 1. Liste des 23 visèmes avec leur symbole et leur contenu (visuo)phonétique majeur.

Notation: * = {i} or {a}; V (c) = {[cVc] / V ∈ {i, a, y}}; C (v) = {[vCv] / v ∈ {i, a, y}};
C (v*) = {[vCi], [vCa] / v ∈ {i, a, y}}; C (*v) = {[iCv], [aCv] / v ∈ {i, a, y}}.

N°	Symbole	[v_iV$_i$$v_i$z]	[$v_j$$c_kV_i$$c_k$$v_j$z]	[$v_i$$c_k$$v_jC_k$$v_i$z]
1	{a}	a	a (ʀ, l)	l (a), ʀ (a*, *a)
2	{i}	i	a (z)	
3	{y}	y, u, ø, o, ɔ	y (b, z, ʒ, ʀ, l)	[z, ʒ, ʀ, l] (y)
4	{e}	e, ɛ, ɛ̃	i (ʀ, l)	ʀ (i), l (i*, *i)
5	{ɔ}	ɔ		
6	{œ}	œ, œ̃		
7	{ã}	ã		
8	{ab}		a (b, v)	
9	{aʒ}		a (ʒ)	
10	{ib}		i (b, v)	
11	{iz}		i (z)	z (**)
12	{iʒ}		i (ʒ)	ʒ (**)
13	{yv}		y (v)	v (y)
14	{b*}			b (**)
15	{by}			b (y*, *y)
16	{v*}			v (**)
17	{zy}			[v, z] (y*, *y)
18	{ʀy}			ʀ (y*, *y)
19	{ly}			l (y*, *y)
20	{ʒy}			ʒ (y*, *y)
21	{yby}			b (y)
22	{#}	forme préphonatoire (lèvres entrouvertes)		
23	{=}	forme de repos (lèvres fermées)		

3. ARCHITECTURE DU SYSTÈME

3.a Structure générale du système

Le système que nous avons mis au point réalise une synthèse audio-visuelle de la parole à partir du texte par concaténation d'unités (acoustiques et optiques) préstockées. Comme le montre la figure 2., il utilise un dictionnaire d'environ 2000 polysons (avec sa table d'accès) sur lesquels sont réalisées des modifications prosodiques (par traitement PSOLA) d'une part, et une table d'accès à un dictionnaire de 23 visèmes nécessaire pour l'affichage final, d'autre part. Plusieurs dictionnaires peuvent être utilisés indifféremment (voix ou visage d'homme, de femme, etc.). L'entrée du système est un texte orthographique français quelconque ; les sorties sont un fichier d'échantillons audio sur 16 bits à 16 kHz et un fichier ASCII descripteur de la séquence d'images correspondante à 25 Hz.

L'ensemble des règles transformationelles nécessaires à la synthèse acoustique et visuelle à partir du texte (transcriptions orthographique-phonétique, phonétique-visémique, etc.), et des modélisations prosodiques ont été écrites sous le formalisme COMPOST (Bailly & Tran, 1989). Ce COMpilateur Phonétique sur Ordinateur pour la Synthèse de Texte permet la manipulation d'objets par des règles contextuelles. Ces objets sont organisés dans une structure arborescente dont ils sont les feuilles. Ils sont des instances d'objets génériques qui sont regroupés dans leur classe naturelle (mot, syllabe, phonème, archi-visème, visème, image, etc.) afin de les doter d'attributs numériques spécifiques.

Trois grammaires de règles ont été élaborées pour la conversion de l'arbre des phonèmes munis de leur durée (modélisée par le module prosodique de la synthèse acoustique) en une séquence d'images alignées sur le signal acoustique. Elles sont présentées ci-dessous.

3.b Correspondance phonème / archi-visème

Il s'agit d'une réécriture élémentaire. Par *archi-visème*, nous entendons ici un ensemble de configurations anatomiques identiques dans la réalisation d'un ou plusieurs phonèmes ne différant que par des traits articulatoires non visibles (voisement, nasalité, lieu de constriction dans le conduit vocal), et ce indépendamment du contexte. Ainsi, les occlusives bilabiales [p], [b], [m] se regroupent naturellement dans un seul archi-visème (noté V_P dans le tableau 2). La transcription se fait donc terme-à-terme, sans prise en compte du contexte, au seul niveau symbolique. L'archi-visème hérite de la durée du phonème correspondant. Les 14 archi-visèmes que nous utilisons actuellement pour le français sont présentés dans le tableau 2.

Tableau 2. Tableau de correspondance symbolique archi-visème – phonème.

archi-visème	phonèmes correspondants
V_A	a
V_E	e, ɛ, ɛ̃
V_I	i, j
V_U	y, u, ø, o, õ, w, ɥ
V_Œ	œ, œ̃
V_Ã	ã
V_Ω	ɔ
V_B	p, b, m
V_V	f, v
V_Z	s, z
V_J	ʃ, ʒ
V_R	ʀ
V_L	l
V_?	t, d, n, k, g, ɲ, ŋ

3.c Transcription archi-visème / visème

Pour nous, un visème est le code symbolique d'une classe de formes maxillo-labiales similaires regroupant divers allophones. Dans leur étude de nombreuses réalisations contextuelles, Benoît et al. (1992) ont identifié l'existence de 21 visèmes répondant à cette définition dans le jeu articulatoire d'un locuteur français représentatif du "français standard" prononcé sans expression faciale particulière, auxquels ont été ajoutés deux visèmes extra-linguistiques : la forme *préphonatoire*, aux lèvres entrouvertes ; et la forme *de repos*, aux lèvres jointes. La transcription archi-visème / visème utilise donc une grammaire de réécriture prenant en compte les contextes gauche et droit de l'archi-visème. Le visème transcrit hérite des propriétés de durée de l'archi-visème correspondant. Une séquence de visèmes est toujours initialisée par le visème *préphonatoire*, et conclue par le visème *de repos*. Un certain nombre de ces règles sont directement lisibles dans le tableau 1. Ainsi, l'archi-visème V_A entouré de deux archi-visèmes V_Z se transcrit sous la forme du visème noté {i} (cf. ligne 2, col. 4 du tableau 1), ou encore, l'archi-visème V_I dans un contexte V_R ou V_L se transcrit sous la forme du visème noté {e} (ligne 4, col. 4 ibidem). Nous soulignons par ces deux exemples la forte influence du contexte sur la forme maxillo-labiale d'une réalisation allophonique...

Toute la combinatoire contextuelle possible sur l'ensemble des archi-visèmes n'est pas implémentée dans la version actuelle du système présenté : elle est encore à l'étude.

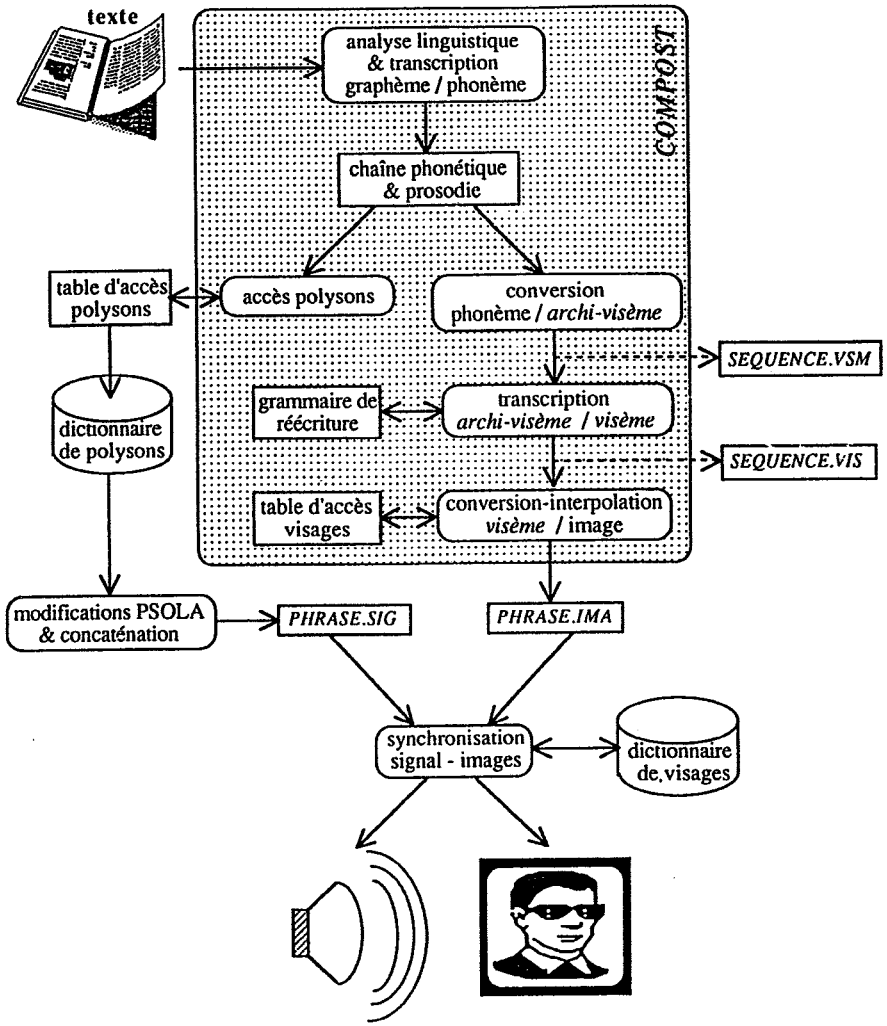


Figure 2. Architecture du synthétiseur audio-visuel à partir du texte de l'ICP. L'environnement COMPOST est figuré dans le cadre pointillé. Le système fonctionne sur un PC doté de cartes OROS AU21 et VGA.

3.d Transcription visème / image

Le dictionnaire d'images faciales utilisé par la synthèse doit naturellement être étiqueté de façon à ce que chaque image corresponde à (au moins) l'un des visèmes obtenus précédemment. Différents dictionnaires peuvent évidemment être utilisés : soit par simple modification de caractéristiques indépendantes des mouvements de la parole (nez, cheveux, etc.), soit par différents angles de vue (face, profil, etc.), soit enfin par la constitution d'un nouveau dictionnaire de formes avec un autre locuteur, à l'instar des dictionnaires d'unités acoustiques.

Cette dernière grammaire ne se limite pas au niveau symbolique. La transcription d'une chaîne de visèmes avec leur durée en séquence d'images étiquetées dans le dictionnaire utilisé doit tenir compte non seulement de la fréquence d'affichage des images sur l'écran (25 Hz dans notre cas pour des images 172 x 103 pixels), mais aussi de la possibilité d'interpolation spatiale d'une image entre deux images-cibles successives si la durée qui les sépare le permet. En effet, Le fichier de sortie contient une séquence d'images numérotées devant être visualisées à périodicité fixe par le synchronisateur son / image. Cette étape consiste donc à échantillonner les images nécessaires, en fonction de la durée prévue du visème courant et du décalage incrémental introduit par l'affichage des visèmes précédents (de durée constante). En outre, une matrice des interpolations possibles entre visèmes permet à ce stade d'intercaler, si cela est possible, l'image correspondant à un troisième visème entre deux (ou plus) images différentes. En plus clair, et sur l'exemple d'une transition entre les visèmes {i} et {a}, il est anatomiquement légitime d'intercaler l'image correspondant au visème {e} (voir figure 1.b) si la durée entre les images de {i} et de {a} est supérieure à 40 ms.

3.e Synchronisation image / son

Les deux étages de sortie du système de synthèse audio-visuelle à partir du texte fournissent un fichier acoustique binaire échantillonné à 16 kHz (PHRASE.SIG) d'une part, et un fichier ASCII descripteur de la séquence d'images à afficher à 25 Hz (PHRASE.IMA) d'autre part. Un programme *ad hoc* lit ces deux fichiers et les convertit de façon synchrone en signal audio via la carte OROS AU21 et en images animées sur la carte VGA respectivement. La commande d'affichage d'une nouvelle image est effectuée entre deux transferts du disque vers la mémoire du CNA, pendant le bref instant où l'unité centrale est disponible.

Un délai d'offset entre les deux canaux d'information image et son a été ajusté expérimentalement (à la trame près d'un enregistrement vidéo 50 Hz) de façon à ce que la première image "parole" soit synchronisée avec le début du signal audio. Ce retard est mis à profit pour afficher l'image correspondant à la forme préphonatoire {#}. En fin de séquence, la position de repos *lèvres fermées* reste affichée quelques instants.

6. CONCLUSION

Cet outil de synthèse de visage parlant à partir du texte avait été élaboré par Woodward en 1991 dans une première version. Il en est aujourd'hui à sa deuxième version et reste donc dans un état encore primitif.! Cependant, il a été conçu de façon à être évolutif. Il sera prochainement amélioré par une adjonction et une optimisation de règles de transcription. En outre, il est actuellement implanté sur un PC 386 ; et sa migration sur une machine plus puissante permettrait une visualisation de visages, soit plus grands et plus détaillés, soit plus fréquents. Enfin, l'intervention d'un infographiste permettrait de remplacer les rares et pauvres contours de notre visage artificiel par un dessin plus naturel – ou caricatural – qui lui fait actuellement défaut d'esthétique...

Mais la toute prochaine étape de notre travail consistera à quantifier l'apport d'intelligibilité de cette animation faciale à la parole synthétique par un test perceptif analogue à celui qui vient d'être réalisé avec de la parole et un visage naturels, et dont les résultats sont présentés par Mohamadi et Benoît dans ces mêmes Journées d'Etude sur la Parole.

Remerciements

Ce travail a été partiellement financé par l'ACCT. Nous remercions J. Zeiliger, A. Arnal, T. Lallouache et C. Abry pour leur précieuse collaboration.

Références

Bailly, G., & Tran, A. (1989), "Compost: A rule-compiler for speech synthesis", *Proc. 1st Eurospeech Conf.*, Paris, France, 136-139.

Benoit, C., Boë, L.J. & Abry, C.(1991), "The effect of context on labiality in French", *Proc. 2nd Eurospeech Conf.*, Gênes, Italie, 153-156.

Benoît, C., Lallouache, T., Mohamadi, T. & Abry, C.(1992), "A set of French visemes for visual Speech synthesis", in *Talking Machines : Theories, Models and Applications*, Bailly & Benoit Eds., Elsevier, North Holland, 485-504.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974), "Auditory and visual contribution to the perception of consonants", *J. Speech & Hear. Res.*, 17, 608-618.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of speech stimuli", *J. Speech & Hear. Res.*, 12, 423-425.

Erber, N. P. (1975), "Auditory-visual perception of speech", *J. Speech & Hear. Dis.*, 40, 481-492.

Mohamadi, T., & Benoît, C. (1992), "Le gain des lèvres : intelligibilité auditive et visuelle de la parole bruitée en français", ce volume.

Neely, K. K. (1956), "Effect of visual factors on the intelligibility of speech", *J. Acoust. Soc. Am.*, 28, 1275-1277.

Sumby, W.H. & Pollack, I. (1954), "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26, 212-215.

Summerfield, Q. (1979), "Use of visual information for phonetic perception". *Phonetica*, 36, 314-331.

Woodward, P. (1991), *Synthèse de visage parlant*, Mémoire de DEA Signal-Image-Parole, ENSERG, Grenoble, 43 p.

LE GAIN DES LÈVRES : INTELLIGIBILITÉ AUDITIVE ET VISUELLE DE LA PAROLE BRUITÉE EN FRANÇAIS

T. MOHAMADI & C. BENOIT

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N° 368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

La perception bimodale permet une meilleure compréhension de la parole que sa seule audition. Dans cet article, nous quantifions l'intelligibilité apportée par le visage du locuteur à la perception de stimuli audio, en fonction de leur dégradation par du bruit blanc. Dix-huit sujets doués d'une audition et d'une vision correctes ont été soumis à un test d'identification à choix fermé portant sur trois voyelles et six consonnes du français, en perception auditive seule d'une part, et audio-visuelle d'autre part. Les taux moyens d'identification correcte fournissent une première mesure de l'amélioration globale obtenue en perception bimodale. La comparaison des matrices de confusion permet en outre de relativiser les effets respectifs de chacune des voyelles et des consonnes. Enfin, nous mesurons l'importance de l'influence contextuelle des trois voyelles [a, i, y] sur les intelligibilités auditive et audio-visuelle des six consonnes [b, v, z, ʒ, ʁ, l] de notre corpus.

1. INTRODUCTION

Avoir la possibilité d'observer le visage d'un locuteur est un avantage considérable pour la compréhension du discours, et tout particulièrement dans un milieu bruité ou réverbérant. Plusieurs travaux ont visé à quantifier l'apport de la vision dans la reconnaissance de la parole dégradée (Sumby & Pollack, 1954; Erber, 1969; Summerfield, 1979). Ces études ont montré que, même pour des sujets normalement entendants et non entraînés à la lecture labiale, les informations faciales augmentent notablement l'intelligibilité de la parole dégradée.

Dans le bruit, les deux modalités visuelle et auditive se complètent lors de la perception de la parole. En effet, ce qui a été dégradé par le bruit dans le spectre de la parole peut être récupéré par la vision des aspects les plus importants de la configuration des lèvres, des dents, et de la langue, qui déterminent le lieu d'articulation de nombreuses consonnes (McGrath et al., 1984).

Sumby et Pollack (1954), puis Erber (1969), ont mesuré le gain en intelligibilité de la perception audio-visuelle (AV) sur l'audition seule (A), en fonction du bruit (S/B) ajouté au signal acoustique. Pour les premiers, ce gain dépend aussi du nombre de mots à identifier : le taux d'intelligibilité de la parole décroît lorsque S/B décroît et/ou le nombre de stimuli augmente, et ce quelle que soit la ou les modalité(s). L'apport de la vision à l'intelligibilité de la parole augmente avec la dégradation acoustique. Pour Erber, cet apport reste stable quand S/B < –30 dB, soit une simple lecture labiale, mais, à –12 dB, le taux de reconnaissance passe de 20% en A à 80% en AV.

MacLeod et Summerfield (1986) ont quantifié l'apport de la perception visuelle comme la différence, en S/B (i.e., en dB), entre les seuils de perception de la parole en A et en AV. Ils ont ainsi mesuré un bénéfice moyen de 11 dB apporté par la lecture faciale.

À ce jour, toutes les expériences connues ont été menées sur la langue anglaise. Nous avons étudié ici le français, en étendant l'analyse de nos résultats à une étude des effets contextuels voyelle-consonne. Ces résultats serviront en outre de référence à des expériences ultérieures pour quantifier et comparer les intelligibilités A et AV de visages parlants synthétiques en cours de développement dans notre laboratoire.

2. MÉTHODE

2.a Corpus

A partir d'un corpus utilisé pour une étude analytique (Benoît et al., 1992), nous avons sélectionné des stimuli de la forme [$V_iC_jV_iC_jV_iz$] (e.g., [iviviz], ou [ababaz], etc.), portés par la phrase interrogative : C'est pas "$V_iC_jV_iC_jV_iz$" ? Notre choix s'est limité à V_i = [i, a, y] et à C_j = [b, v, z, ʒ, ʁ, l]. Les trois voyelles correspondent aux positions extrêmes du mouvement labial des voyelles ; les quatre premières consonnes sont supposées avoir un effet majeur aux lèvres ; les deux dernières sont supposées neutres, ou jouer un rôle encore mal connu en ce qui concerne la labialité. Dix-huit stimuli différents constituaient donc au total notre test.

2.b stimuli et préparation du test

L'enregistrement et la prise de vue ont été réalisés dans une chambre sourde. Le locuteur, fortement éclairé, a été filmé de face et de profil. Seule, la vue de face, en noir et blanc a été utilisée pour ce test. Les stimuli sélectionnés ont été copiés automatiquement sur une bande vidéo, à l'aide d'un poste de travail spécialisé (Lallouache, 1991) pilotant deux magnétoscopes, l'un en lecture, le second en enregistrement. L'audio a été enregistré sur un seul canal, le deuxième canal étant réservé à l'enregistrement du bruit.

Le bruit de masquage a été produit par un générateur de bruit blanc (spectre plat entre 20 Hz et 20 kHz). Sur le magnétoscope enregistreur, les deux niveaux d'enregistrement audio (canal 1 pour le bruit, canal 2 pour la parole) ont été réglés à l'aide du vumètre, à 0 dB pour le bruit et à –36 dB pour le signal de la parole (la réalisation du phonème [a] de "c'est pas" dans la phrase porteuse servant de référence). Le dispositif expérimental est présenté sur la Figure 1.

Chaque stimulus, son mode de présentation (0 pour A et 1 pour AV), les numéros d'images de début et de fin de sa phrase porteuse, ainsi que le gain de sortie du bruit, ont été préalablement stockés, sur une même ligne dans un fichier descripteur ASCII servant de scénario au dispositif expérimental (voir Figure 2).

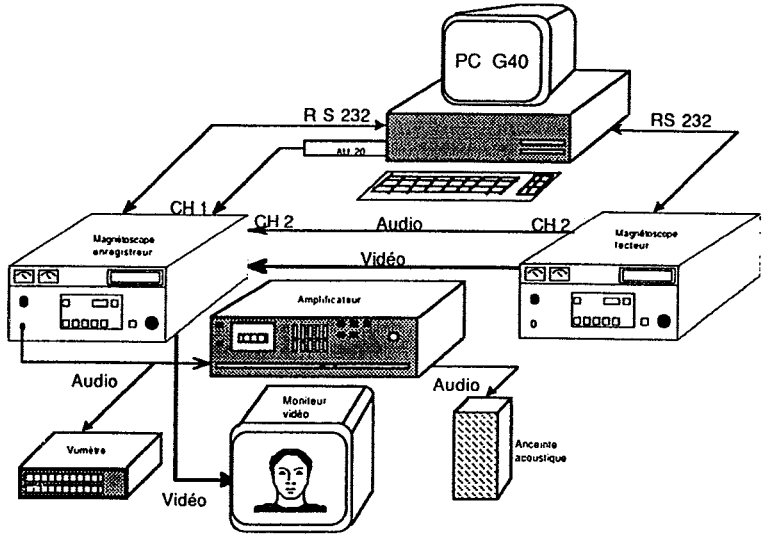


Figure 1. Dispositif expérimental pour la présentation automatique des stimuli

Stimulus	Mode A ou AV	trame Début	trame Fin	S/B
IBIBI	1	6468	6534	-36
UVUVU	1	6408	6468	-30
ILILI	1	7984	8040	-24
...	...	...	...	...
ABABA	1	5305	5369	-24

Figure 2. Exemple de fichier descripteur (partiel) du test en mode AV

2.c Les sujets

Nous avons retenu 18 sujets français, des deux sexes (11 femmes et 7 hommes), d'âge compris entre 19 et 26 ans (21,5 en moyenne), sans formation particulière en sciences de la parole ni familiarité avec des malentendants. Un audiogramme et un contrôle de l'acuité visuelle de chacun des candidats ont été effectués au début de chaque session. Un seul candidat a été rejeté, pour déficience auditive. Les sujets ont été rémunérés pour leur participation.

2.d Procédure

Le test a été divisé en deux séances individuelles (A et AV) de 24 minutes en chambre sourde. L'ordre de présentation était différent d'un sujet à l'autre (décalage de 5 stimuli à chaque session), avec le même sens de dégradation (du moins au plus dégradé). L'ordre des deux séances a été contrebalancé d'un sujet à l'autre. En face, et à 1,5 m du sujet assis sur une chaise, étaient disposés un moniteur vidéo et une enceinte acoustique.

Les sujets ont reçu des directives, puis ont subi une phase de familiarisation, avec présentation de 5 stimuli, avant chaque séance. Les réponses correspondantes n'ont pas été prises en compte. Cinq feuilles de réponse avec les six consonnes et les trois voyelles disposées sur la même ligne (voir Figure 3) ont été fournies au sujet à chaque séance. Il lui était demandé d'entourer la consonne et la voyelle perçues à chaque stimulus, ou de barrer la ligne correspondante dans les cas d'imperceptibilité (pour éviter l'oubli ou le saut de ligne). Une voix enregistrée invitait le sujet à changer de feuille de réponse tous les 18 stimuli. Un bip sonore précédait de 2 s chaque stimulus. Un délai de 15 s après chaque stimulus (mis à profit pour la recherche automatique de début du stimulus suivant sur la bande) offrait au sujet un temps de réponse confortable.

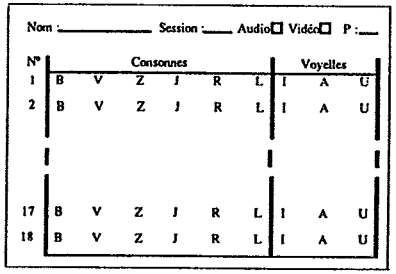


Figure 3. Vue partielle d'une feuille de réponse

3. SCORES D'INTELLIGIBILITÉ

3.a Résultats globaux

Les réponses de chaque sujet ont été comparées au fichier de référence correspondant. Dans un premier temps, la réponse a été notée fausse en cas d'erreur sur la consonne et/ou sur la voyelle, ou en cas de non-réponse. Le nombre de réponses correctes sur l'ensemble des dix-huit stimuli et des dix-huit sujets, pour chacun des rapports signal à bruit, a été converti en pourcentage. Pour chaque niveau de bruit, l'écart-type a été calculé sur l'ensemble des dix-huit sujets. Les valeurs calculées, ainsi que les deux courbes qui les interpolent sont présentés à la Figure 4.

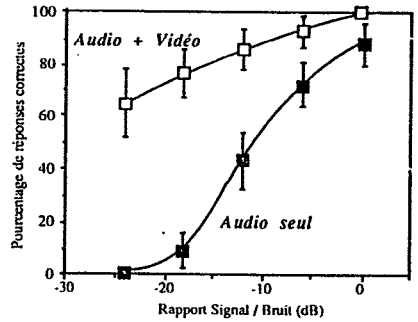


Figure 4. Taux moyens d'identification de 18 stimuli, en présentation audio seule (courbe du bas), et en audio-visuel (courbe du haut), par 18 sujets, en fonction de la dégradation audio. Les écarts-type des réponses individuelles sont figurés par les segments de droite verticaux sur chaque point de mesure. Les courbes joignant les 5 points ont été obtenues par interpolation.

Le taux d'identification correcte en audio seul chute brusquement de 72% à 8% pour un écart de rapport signal à bruit de 12 dB (entre S/B = –6 dB et S/B = –18 dB), ce qui est comparable (aux conditions de mesures près) aux observations d'Erber (1969) pour l'anglais. En AV, le taux diminue, pour le même écart, de 93% à 77%. A –24 dB le taux en A est presque nul (0,62%, quand l'aléatoire se situerait à 5,6%), mais il atteint 65% en AV. Ce résultat spectaculaire confirme la facilité avec laquelle les sujets ont pu identifier les stimuli en utilisant quasi-exclusivement la lecture labiale. Nos observations concordent avec celles de Sumby et Pollack (1954) pour le même nombre de mots à identifier : 75% à –30 dB pour 16 mots (à un décalage près d'environ 10 dB en S/B entre nos mesures et les leurs, dû essentiellement au choix de la référence "signal" ; nous y reviendrons plus bas).

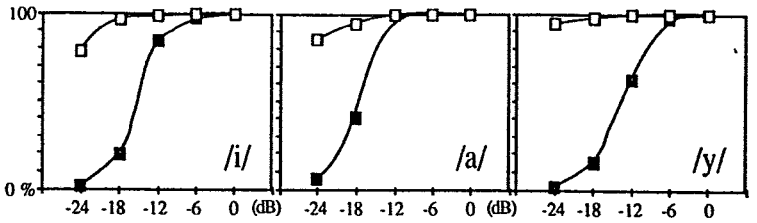


Figure 5. Intelligibilité A (carrés noirs) et AV (carrés blancs) de [i, a, y], esxprimée en % de voyelles correctement identifiées, tous contextes consonantiques confondus, par 18 sujets.

3.b Intelligibilité contextuelle

3.2.1 Intelligibilité comparée des voyelles [i, a, y]

Pour permettre une lecture plus détaillée de l'intelligibilité phonétique, la Figure 5 présente les courbes du taux moyen d'intelligibilité de chacune des trois voyelles [i, a, y], tous contextes consonantiques confondus, en fonction du S/B dans les deux conditions de présentation, A et AV.

En perception auditive seule, [a] est plus intelligible que [i], lui-même plus intelligible que [y], toutes choses égales par ailleurs (i.e., tous auditeurs et contextes confondus, cf. à S/B = -18 et -12 dB). Cette observation se justifie par l'intensité intrinsèque des trois voyelles testées : qu'il s'agisse de l'anglais américain, pour lequel Black (1949) a observé une intensité moyenne du [ɑ] de 3,69 dB supérieure à celle du [i], ou du français, pour lequel Guérin et Boë (1978), entre autres, ont observé la même différence entre [a] et [i], l'intensité moyenne de [y] étant comparable à celle du [i]. Or, le S/B calculé ici repose sur l'intensité "signal" de la voyelle [a] de référence mesurée dans "c'est pas...", et non sur l'intensité moyenne du stimulus, ou de la voyelle test. Il faut donc relativiser cette différence d'intelligibilité des trois voyelles par rapport à leur intensité intrinsèque. En effet, les courbes d'intelligibilité acoustique du [a] correspondent grossièrement à celles du [i] ou du [y] à un glissement près de 3 dB environ le long de l'axe des abscisses.

En perception bimodale et sous forte dégradation acoustique (S/B ≤ -18 dB), [y] est mieux identifié que [a], lequel est mieux identifié que [i]. Il faut certainement invoquer ici la spécificité du jeu maxillo-labial nécessaire à la production de ces trois voyelles, et sa robustesse relative aux modifications engendrées par l'environnement consonantique.

3.2.2 Effet du contexte vocalique sur l'intelligibilité des consonnes

L'effet global des trois voyelles sur les intelligibilités A et AV des six consonnes testées est présenté Figure 6, par la projection sur deux axes (intelligibilité A, intelligibilité AV) des différents points obtenus en moyennant sur les cinq conditions de dégradation acoustique les scores de chacune des consonnes, et ce pour chaque contexte vocalique. Chaque consonne est représentée graphiquement par son pourcentage moyen d'identification correcte, tous contextes confondus. De celui-ci partent trois flèches désignant le score obtenu dans chacun des trois contextes vocaliques. Un important effet contextuel est ainsi mis en relief : le contexte vocalique [a] est celui qui favorise nettement l'intelligibilité de toutes les consonnes, tant en A qu'en AV. À l'inverse, [y] est le contexte pessimisant, et ce pour les deux types de présentation (A passe de 30 à 65% et AV de 55 à 90% selon que [z] est présenté en contexte [y] ou [a], par exemple). En revanche, le contexte [i] ne laisse pas apparaître de loi systématique quant à son influence sur l'intelligibilité des consonnes : cette influence est faible, et sa tendance serait à un gain en AV par rapport à la moyenne des autres contextes.

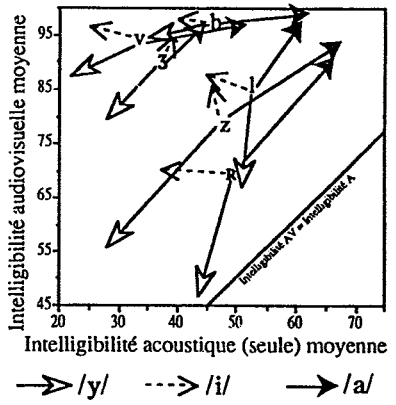


Figure 6. Effet des voyelles [i], [a], et [y] sur l'intelligibilité des consonnes [b, v, z, ʒ, ʁ, l] exprimée en % de consonnes correctement identifiées (voir texte).

Si l'on essaye d'extrapoler ces observations à un gain d'intelligibilité apporté par la seule vision du locuteur, et à supposer en première approximation que les deux modalités sont indépendantes l'une de l'autre, il est intéressant de considérer cette projection de la Figure 6 par rapport à la droite {Score AV - Score A = 0} représentée en bas à droite du plan. Tous les points expérimentaux sont situés au-dessus de cette droite, confirmant ainsi le gain d'intelligibilité apporté par la lecture labiale. En revanche, le remarquable effet du contexte [a] vs. [y] a tendance à se situer parallèlement à cette droite, laissant supposer que le gain de la lecture labiale est quasi-indépendant du contexte vocalique pour ce qui concerne l'intelligibilité des consonnes, exceptions faites de [ʀ] et de [l] en contexte [y]. C'est le contexte [i] qui optimise ici l'apport de la lecture labiale des consonnes, puisque les six flèches matérialisant son effet sur la Figure 6 s'éloignent de la droite A = AV.

Il est permis de supposer que l'intensité intrinsèque des voyelles influe sur celle des consonnes environnantes, pour justifier la contribution du contexte [a] à l'apport d'intelligibilité en audio seul. De la même façon, il faut sans doute invoquer la forte coarticulation de [y] (voyelle la plus intelligible en AV) sur les consonnes qui le jouxtent, et la relative plasticité de la forme du [i] (voyelle la moins intelligible en AV) qui dépend souvent davantage du contexte consonantique que l'inverse, pour justifier les différences de gain moyen entre A et AV selon que les consonnes sont perçues en contexte [y] (gain moyen AV – A faible) et [i] (gain moyen AV – A élevé).

4. ANALYSE DES CONFUSIONS

Toutes les réponses des sujets ont été saisies dans une base de données qui nous a servi à établir des matrices de confusion entre les trois voyelles [i, a, y] d'une part, et les six consonnes [b, v, z, ʒ, ʀ, l] d'autre part. Nous avons retenu la condition S/B = –12 dB afin de comparer au mieux les deux modes de perception.

4.a Confusion des voyelles à –12 dB

La Figure 7 présente les deux matrices de confusion des voyelles [i, a, y], à –12 dB, par nos dix-huit sujets, les six contextes consonantiques confondus. Le point d'interrogation figure les non-réponses.

En condition A, la voyelle [a] n'est confondue avec aucune autre. Par contre les voyelles [i] et [y] sont partiellement confondues entre elles, la voyelle [i] ayant tendance à un effet attracteur des réponses (125 percepts pour 108 stimuli) par rapport aux réponses sur la voyelle [y] (79 percepts seulement).

Une désambiguïsation très nette est réalisée en mode AV où les trois voyelles sont perçues correctement quasi-systématiquement.

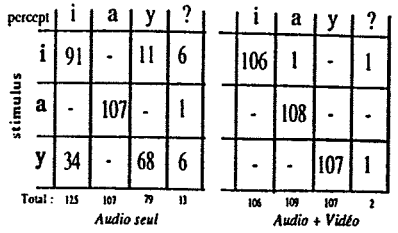


Figure 7. Matrices de confusion des voyelles [i, a, y], tous contextes consonantiques confondus, par 18 sujets, à S/B = –12 dB. A gauche, confusions auditives seules. A droite, confusions audio-visuelles.

4.b Confusion des consonnes à -12 dB

La Figure 8 présente les matrices de confusion des six consonnes testées dans les deux conditions A et AV à un niveau de dégradation S/B = –12 dB.

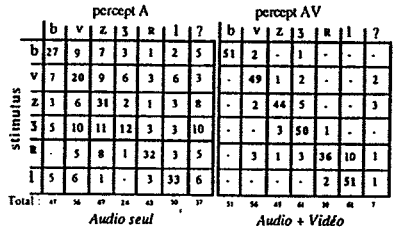


Figure 8. Matrices de confusion des consonnes [b, v, z, ʒ, ʀ, l], tous contextes vocaliques confondus, par 18 sujets, à S/B = –12 dB. A gauche, confusions auditives seules. A droite, confusions audio-visuelles.

Les consonnes [l, ʀ, z] sont les moins confondues en perception auditive seule. Le score le plus faible est celui de la consonne [ʒ] qui est fréquemment identifiée comme [z] ou [v]. Cet effet n'est pas réciproque : [ʒ] est rarement proposé comme réponse (24 percepts pour 54 stimuli), quand [z] offre, lui, une réponse attractive (67 percepts). Il est intéressant de noter ici la faible confusion [ʀ] / [l] en présentation A.

En présentation AV, la désambiguïsation est encore ici très marquée. Toutes les confusions deux à deux sont en nette diminution, hormis le cas particulier du [ʀ] pour lequel les réponses [ʒ], et surtout [l], sont en augmentation ! En effet, 10 fois sur 54 (au lieu de 3 en mode A), [ʀ] est confondu avec [l]. Comme nous l'avons souligné plus haut, c'est le contexte [y] qui est majoritairement responsable de cette exceptionnelle dégradation de l'intelligibilité au passage de A à AV. La forme maxillo-labiale du [ʀ] et du [l] étant très variable,

la vision du contexte protrus du [y] fait augmenter la confusion [ʀ] -> [l]. Nous n'avons pas d'explication à proposer pour ce phénomène d'attraction non réciproque entre "liquides aux lèvres", sauf à supposer que la réalisation /yʀyʀyz/ testée présente des caractéristiques visuelles particulièrement proches de celles d'un /ylylyz/. Cette interprétation est en accord avec des mesures géométriques antérieures (Benoît et al., 1992), portant sur plusieurs répétitions de /yʀy/ et de /yly/, et qui permettent tout aussi bien d'envisager l'effet attracteur inverse [l] -> [ʀ] avec d'autres réalisations.

4.c Confusions visuelles

La Figure 9 présente les deux matrices de confusion des voyelles (à gauche) et des consonnes (à droite), en présentation AV, à un rapport signal à bruit de -24 dB où aucun son n'est perçu dans le bruit (i.e., en présentation visuelle seule en première approximation).

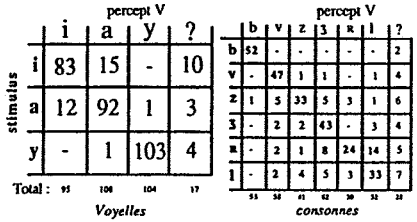


Figure 9. Matrices de confusion des voyelles [i, a, y], tous contextes consonantiques confondus (à gauche) et des consonnes [b, v, z, ʒ, ʀ, l], tous contextes vocaliques confondus (à droite), par 18 sujets, à S/B = –24 dB, i.e. pratiquement en mode "video seul".

Presque toute l'information est apportée ici par la vision du visage du locuteur, en présence d'un bruit de niveau très élevé. On constate que ce sont la voyelle [y] et la consonne [b] qui présentent les formes les plus discriminables. C'est la grande précision du mouvement labial nécessaire à leur production qui est responsable de leur facilité d'identification, quel que soit le contexte : si l'occlusion totale pour le [b] est un truisme, il est bon de rappeler que l'aire intérolabiale du [y] est toujours inférieure à 100 mm^2 (Abry & Boë, 1986). A l'opposé, [i] et [ʀ] sont les moins identifiables, car les plus "élastiques" de notre corpus.

5. CONCLUSION

Les résultats de notre étude pour le français rejoignent globalement ceux obtenus en anglais. Leur analyse détaillée est, par contre, spécifique de la langue : le [ʀ] français a phonétiquement très peu de rapport avec son homologue [ɹ] anglais, par exemple. L'effet du contexte vocalique sur l'intelligibilité des consonnes en parole dégradée, quel que soit le mode de présentation, s'il est confirmé avec d'autres locuteurs français, devra être étudié avec profit par les orthophonistes et les enseignants de français.

Cette étude démontre le remarquable gain d'intelligibilité apporté à la parole acoustiquement dégradée par la présentation du visage d'un locuteur : elle justifie donc l'intérêt technologique d'un développement de synthétiseurs audio-visuels de la parole à partir du texte, tels que celui présenté à ces mêmes Journées d'Etude (Woodward et al., 1992).

Remerciements

Ce travail a été particllement financé par l'ACCT. Grand merci à Jérôme Zeiliger, Marie-Agnès Cathiard, et Alain Arnal pour leur support technique, ainsi qu'à Christian Abry, Louis-Jean Boë, et Tahar Lallouache pour leurs conseils et leur intérêt pour cette étude...

Références

Abry, C. & **Boë**, L.J. (1986), "Laws for lips", *Speech Communication*, 5, 97-104.

Benoît, C., **Lallouache**, T., **Mohamadi**, T. & **Abry**, C. (1992), "A set of French visemes for visual Speech synthesis", in *Talking Machines : Theories, Models and Applications*, G. Bailly & C. Benoit Eds., Elsevier, North Holland.

Black, J. W. (1949), "Natural frequency, duration, and intensity of vowels in reading", *J. Speech & Hear. Disorders.*, 14, 216-221.

Cathiard, M.A. (1988/1989), "La perception visuelle de la parole: Aperçu de l'état des connaissances". *Bull. Inst. Phonétique Grenoble*, 17-18, 109-193.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of speech stimuli", *J. Speech & Hear. Res.*, 12, 423-425.

Guérin, B. & **Boë**, L.J. (1978), "Etude d'un indice acoustique des voyelles: la puissance intrinsèque", *9° JEP*, Lannion, 1, 167-176.

Lallouache, T. (1991), "Un poste 'visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres", *Thèse de Doctorat*, INP, Grenoble.

MacLeod, A. & **Summerfield**, Q. (1986), "Quantifying the contribution of vision to speech perception in noise", *British J. of Audiology*, 21, 131-141.

Sumby, W.H. & **Pollack**, I. (1954), "Visual contribution to speech intelligibility in noise", *J. Acoust. Soc. Am.*, 26 (2), 212-215.

Summerfield, Q. (1979), "Use of visual information for phonetic perception", *Phonetica*, 36, 314-331.

McGrath, M., **Summerfield**, Q. & **Brooke**, M. (1984), "Roles of lips and teeth in lipreading vowels", *Proc. of the Inst. of Acous.*, 6, 401-408.

Woodward, P., **Mohamadi**, T., **Benoît**, C., & **Bailly**, G. (1992), "Synthèse à partir du texte d'un visage parlant français", *ce volume*.

UN MODÈLE 3D DE LÈVRES PARLANTES

A. ADJOUDANI, T. GUIARD-MARIGNY, B. LE GOFF & C. BENOIT

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N°368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

Un modèle de lèvres a été élaboré à partir d'analyses labiométriques réalisées sur le visage d'un locuteur français. Un premier modèle 2D a d'abord été déterminé en recherchant des équations mathématiques simples ajustant au mieux les contours réels de lèvres vues de face. Les paramètres de commande ont été réduits à trois mesures anatomiques, par analyse de données multidimensionnelles. Un modèle 3D à cinq paramètres de contrôle a ensuite été mis au point et implanté sur une station graphique. Son animation est réalisée à partir de fichiers contenant les cinq paramètres préalablement stockés (50 par seconde) ou bien mesurés simultanément (à 12,5 Hz) sur les lèvres d'un locuteur. Ce modèle a enfin été évalué en termes d'apport d'intelligibilité à la parole naturelle dégradée.

Mots-clés : modèle 3D de lèvres, analyse vidéo

1.INTRODUCTION

Même si la modalité auditive est dominante dans la parole, il a été démontré que la modalité visuelle permet de mieux comprendre la parole, notamment en conditions d'audition défavorables : lorsque l'environnement est bruyant (Sumby & Pollack, 1954 ; Neely, 1956 ; Erber, 1969, 1975 ; Binnie et al., 1974 ; Summerfield, 1979 ; Mohamadi & Benoît, 1992),, lorsque le texte prononcé est linguistiquement complexe ou quand le langage utilisé n'est pas familier (Reisberg et al, 1987). Ainsi, des visages de synthèse ont été développés pour renforcer l'intelligibilité des synthétiseurs de parole qui reste encore bien inférieure à celle des humains.

En 1985, McGrath a montré que les lèvres peuvent rendre compte, à elles seules, des deux tiers de l'intelligibilité visuelle de la parole apportée en moyenne par la vision du visage d'un locuteur. C'est pourquoi un visage de synthèse dédié à la parole doit posséder un modèle perfectionné de lèvres.

Les lèvres, contrairement aux autres régions du visage humain, sont essentiellement caractérisées par leurs contours. L'élaboration du modèle de lèvres présenté ci-dessous a donc essentiellement consisté à rechercher des équations algébriques approximant au mieux les contours réels des lèvres d'un locuteur dans les trois dimensions.

2.LE MODÈLE 2D DE LÈVRES

Dans un premier temps, un modèle de lèvres à 2 dimensions a été élaboré par Guiard-Marigny (1992) à partir des 23 formes maxillo-labiales de référence mises en évidence par Benoît et al. (1992). Ces 23 "visèmes", propre à un locuteur français, permettent de rendre compte d'une grande partie des formes de bouches utilisées pour la parole en français. Guiard-Marigny (1992) a approximé, par un nombre réduit d'équations mathématiques simples, les contours externe et interne, vus de face, des lèvres du locuteur ayant prononcé le corpus de référence. Pour n'utiliser que des fonctions mathématiques simples il a divisé le contour du vermillon en 3 parties, comme le montre la figure 1. Le même type d'équations, polynomial et sinusoïdal, a été utilisé pour le contour externe et le contour interne. Les lèvres du locuteur ont été symétrisées et seule la partie droite de sa bouche a été étudiée.

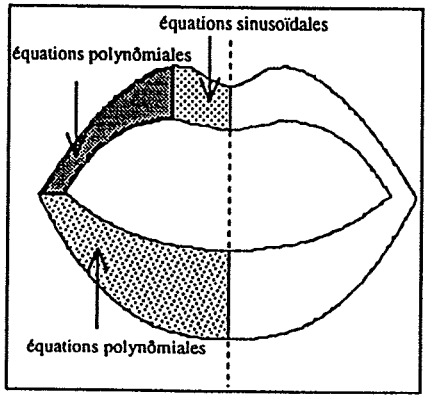


Figure 1 : Les trois parties du modèle de lèvres 2D (Guiard-Marigny, 1992).

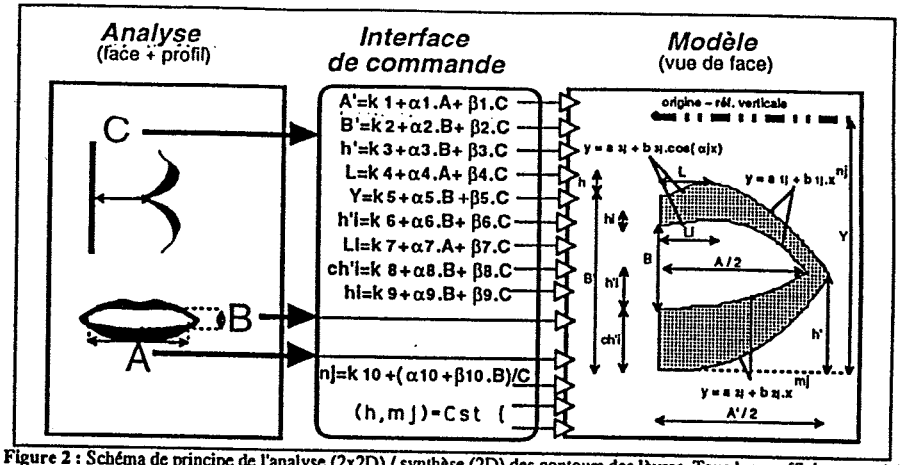


Figure 2 : Schéma de principe de l'analyse (2x2D) / synthèse (2D) des contours des lèvres. Tous les coefficients ont été établis expérimentalement de façon à ajuster au mieux les contours modélisés avec les contours réels (Guiard- Marigny, 1992).

Guiard-Marigny (1992) a déterminé, pour chacun des 23 "visèmes", les valeurs des 15 coefficients intervenant dans les équations des courbes qui ajustent au mieux les contours réels. Ensuite, par analyse de données, il a réduit le nombres de ces paramètres à seulement 3 données anatomiques facilement mesurables sur les images vidéo d'un locuteur : l'étirement horizontal interne (A), la séparation verticale interne (B) et la distance entre une référence verticale sur la vue de profil et le point de contact des lèvres (C). La figure 2 explique en détail le principe d'analyse/synthèse de ce modèle.

3.LE MODÈLE 3D DE LÈVRES

Pour élaborer un modèle 3D à partir du modèle 2D décrit précédemment, Adjoudani (1993) a recherché les équations de la projection des contours labiaux dans un plan perpendiculaire au plan frontal du modèle 2D. La forte influence de la forme de la mandibule sur celle des lèvres l'a conduit à s'intéresser au plan horizontal. Pour donner son volume aux lèvres il a ajouté, sur le modèle de face, 3 contours intermédiaires obtenus par interpolation linéaire. Pour chacun des 5 contours ainsi définis de face, il a déterminé des équations dont les représentations ajustent au mieux la projection du contour dans le plan horizontal. Ainsi, il a obtenu 10 équations de type polynomial (5 pour la lèvre supérieure et 5 pour la lèvre inférieure). Par régressions successives, il a ensuite prédit leurs coefficients à partir des valeurs des paramètres anatomiques mesurés sur le visage du locuteur. La figure 3 montre la représentation en fil de fer du modèle d'Adjoudani. Pour le synthétiser, seuls 2 paramètres anatomiques ont dû être ajoutés aux paramètres de contrôle initiaux du modèle 2D : les protrusions du bord de la lèvre supérieure et du bord de la lèvre inférieure.

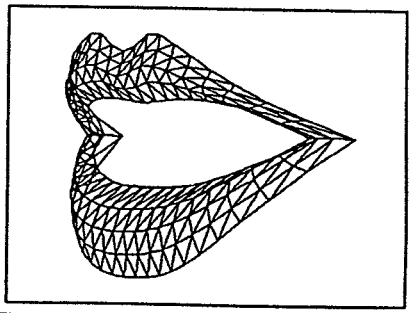


Figure 3 : Représentation en fil de fer du modèle 3D - vue de trois-quarts (Adjoudani, 1993).

4.L'ANIMATION DU MODÈLE

Notre modèle de lèvres 3D a été implanté sur une station graphique (SGI Indigo-ELAN). Pour générer l'image, nous utilisons une surface définie par un réseau de 200 polygones obtenus par échantillonnage des 5 contours. Un lissage de type Gouraud est ensuite appliqué pour atteindre un rendu satisfaisant du vermillon.

Un des systèmes d'animation consiste à analyser préalablement une bande vidéo sur laquelle un locuteur prononce le texte que nous voulons rejouer. Pour cela nous utilisons le poste "vidéo-parole" mis au point par Lallouache (1991). Nous obtenons ainsi un fichier contenant, pour chaque trame vidéo (soit toutes les 20 ms), les valeurs des 5 paramètres de commande de notre modèle. L'animation consiste alors à afficher toutes les 20 ms les lèvres reconstruites à partir des paramètres contenus dans le fichier. La voix du modèle est la voix de notre locuteur, préalablement numérisée, et synchronisée avec l'affichage des images. Ce système permet une animation très naturelle grâce à la forte cadence d'affichage (50 images par seconde) mais surtout grâce à la parfaite synchronisation entre l'image et le son. Cette technique d'animation a ainsi permis de tester l'apport d'intelligibilité de notre modèle en milieu bruité (Le Goff, 1993). L'expérience réalisée est décrite au chapitre 5.

Un système d'animation en temps réel a également été mis au point (Angola, 1993). Une station de travail, munie d'une carte d'acquisition vidéo, analyse puis transmet, à travers une liaison série, les 5 paramètres de commande du modèle toutes les 80 ms (soit 12,5 images par seconde). Un système d'interpolation permet d'obtenir une animation à 25 images par seconde. La voix du locuteur original est retardée d'une durée constante pour compenser le délai dû à l'analyse / synthèse (environ 200 ms sur la plate-forme de travail actuelle), afin de resynchroniser les deux modalités.

5.L'APPORT D'INTELLIGIBILITÉ

Le Goff (1993) a quantifié l'apport d'intelligibilité dû à l'information visuelle synthétique de notre modèle en reconduisant une expérience menée par Benoît et al. (à paraître) sur de la parole naturelle, et en l'étendant à deux conditions visuelles supplémentaires : notre modèle 3D de lèvres seul, et ce modèle superposé au modèle de visage de Parke (1974) modifié par Cohen et Massaro (1990, 1993).

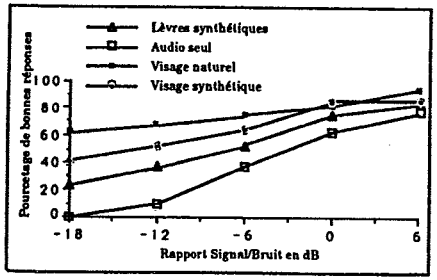


Figure 4 : Courbes représentant les résultats de l'expérience de Le Goff (1993) sur l'apport de l'information visuelle à l'intelligibilité de la parole

Il s'est donc agi, pour 14 sujets, d'identifier, dans chaque condition, une voyelle (V) parmi trois */a/, /i/, /y/* et une consonne (C) parmi six */b/, /v/, /z/, /ʃ/, /R/, /l/*) dans une séquence phonétique VCVCV. Le signal audio correspondant était plus ou moins bruité (S/N de -18 dB à +6 dB). Cette expérience a été menée sur quatre conditions dont trois visuelles : audio seul, audio et visage naturel, audio et lèvres synthétiques, audio et visage synthétique.

Les résultats de Le Goff (1993), représentés sur les figures 4 et 5, confirment globalement ceux de Benoît et al. (à paraître). Quel que soit le niveau de dégradation acoustique, le visage naturel, complet, fournit aux sujets les deux-tiers de l'information manquante; les lèvres synthétiques seules, commandées par cinq paramètres, en fournissent plus du quart; et le visage synthétique complet, contrôlé par six paramètres (les mêmes, plus l'abaissement du menton), en fournit presque la moitié. On le voit, un très faible débit d'information suffit à restituer une grande partie de l'intelligibilité apportée par le visage naturel du locuteur : près de la moitié pour notre modèle de lèvres, et plus de la moitié pour un modèle de visage complet (dans lequel les mouvements de la langue n'étaient cependant pas contrôlés).

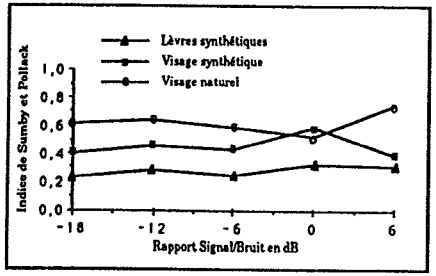


Figure 5 : Proportions de l'intelligibilité acoustique manquante compensée par l'information visuelle (indice défini pour la première fois par Sumby et Pollack (1954) : (I[AV]-I[A])/(1-I[A]).

6.CONCLUSION

Le modèle de lèvres tridimensionnel présenté dans cet article est un modèle de haute définition commandé par seulement cinq paramètres, correspondant à des mesures anatomiques du visage. Il peut ainsi être animé par analyse vidéo. Ses performances ont été mises en évidence par des tests démontrant son apport d'intelligibilité à la parole dégradée. Ce modèle peut donc être avantageusement intégré dans un modèle de visage synthétique dédié à la parole.

Références

Adjoudani, A. (1993), *Élaboration d'un modèle de lèvres 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Angola, O. (1993), *Analyse labiométrique en temps réel par traitement d'images vidéo*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Benoît, C., Lallouache, M.T., **Mohamadi, T. & Abry**, C. (1992), "A set of French visemes for visual speech synthesis", in *Talking Machines: Theories, Models and Designs*, G. Bailly & C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504.

Benoît, C., Mohamadi, T. & Kandell, S.D. (à paraître), "The effect of phonetic context on the the audiovisual intelligibility in French", *Journal of Speech and Hearing Research.*

Binnie, C.A., Montgomery, A.A. & Jackson, P.L. (1974), "Auditory and visual contributions to the perception of consonants", *Journal of Speech & Hearing Research*, 17, 619-630.

Cohen, M.M. & **Massaro**, D.W. (1990), "Synthesis of visible speech", *Behavioral Research Methods and Instrumentation*, 22, 260-263.

Cohen, M.M. & **Massaro**, D.W. (1993), "Modelling coarticulation in synthetic visual speech", Proceedings of *Computer Animation'93*, Magnenat-Thalmann & Thalmann Eds, Genève, Suisse.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of Speech & Hearing Research*, 12, 423-425.

Erber, N.P. (1975), "Auditory-visual perception of speech", *Journal of Speech & Hearing*

Guiard-Marigny, T. (1992), *Animation en temps réel d'un modèle paramétrisé de lèvres*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat de l'Institut National Polytechnique de Grenoble, 214 pp.

Le Goff, B. (1993), *Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

McGrath M. (1985), *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces*, Ph.D. Thesis, Univ. of Nottingham, UK.

Mohamadi, T. & Benoit, C. (1992), "Apport de la vision du locuteur à l'intelligibilité de la parole bruitée", *Bulletin de la Communication Parlée*, 2, Cahiers de l'ICP, INPG, Grenoble, France.

Neely, K.K. (1956), "Effect of visual factors on the intelligibility of speech", *Journal of the Acoustical Society of America*, 28, 1275-1277.

Parke, F.I. (1974), *A parametric model for human faces*, PhD Dissertation, University of Utah, Department of Computer Sciences.

Reisberg, D., McLean, J. & Goldfield A. (1987), "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli", pp. 97-114, in *Hearing by eye: The psychology of lip-reading*, B. Dodd & R. Campbell, Eds, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Sumby, W.H. & Pollack, I. (1954), "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, 26, 212-215.

ANALYSE SYNTHÈSE DE VISAGES PARLANTS

O. ANGOLA, B. LE GOFF, T. GUIARD-MARIGNY, A. ADJOUDANI, C. BENOIT & M. COHEN.

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N°368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Résumé

Nous présentons un système d'analyse et d'animation en temps réel d'un modèle de visage enrichi d'un modèle de lèvres de haute résolution.

L'extraction des paramètres visuels relatifs à la production labiale de la parole ainsi que les modèles tridimensionnels des lèvres et du visage ont été élaborés sur deux ordinateurs couplés de sorte que les lèvres synthétiques seules ou le visage entier peuvent être animés en temps réel ou à partir des données pré-stockées. Les mesures géométriques obtenues à partir des mouvements des lèvres et de la mâchoire d'un locuteur ont été réalisées grâce à des algorithmes de traitement d'images appliqués sur la vue de face et de profil du visage du locuteur. Ainsi, nous arrivons à restituer une grande partie de l'apport de l'intelligibilité de la parole dû à la vision du visage d'un locuteur, avec une faible quantité de paramètres.

Mots-clés : analyse vidéo, visage synthétique parlant, temps réel

1. INTRODUCTION

Un des enjeux actuels des modèles paramétriques de visage est de restituer, avec une faible quantité de données, le maximum de l'information apportée par la vue du visage d'un locuteur. Au Japon, Aizawa et al. (1989) ont élaboré un système d'analyse-synthèse de visages permettant de transmettre à bas débit les différences entre deux images consécutives (à 10 images par seconde) ainsi que les paramètres caractéristiques des expressions faciales basés sur le système de codage FACS (Ekman et Friesen, 1977). En Amérique du Nord, Terzopoulos et Waters (1993) ont utilisé une analyse de contours spécifiques du visage pour déduire les contractions musculaires responsables des déformations de ces contours afin d'animer un modèle paramétrique de visage (Waters, 1987). Plus récemment, une équipe française de l'INA a mis au point un système d'analyse-synthèse de visage en temps réel pour la télé-virtualité (Saulnier et al., 1994). De notre côté, nous avons développé un système d'analyse/synthèse d'un visage parlant dans lequel un nombre réduit de paramètres permet de reconstituer avec une très grande précision les mouvements des lèvres et du menton intervenant dans la production de la parole. Nous présentons, dans un premier temps, le modèle des lèvres et celui du visage que nous utilisons ainsi que les paramètres qui les contrôlent. La technique d'extraction de ces paramètres à partir de l'enregistrement vidéo du visage d'un locuteur est ensuite décrite. Enfin, nous exposons un système d'analyse/synthèse en temps réel dont les performances en terme d'apport d'intelligibilité ouvrent la porte électronique à de nombreuses applications dans le domaine des télécommunications comme dans celui du cinéma d'animation et de synthèse.

2. LES MODÈLES PARAMÉTRIQUES

2.1. LE MODÈLE DE LÈVRES

Le modèle de lèvres 3-D de haute résolution a été élaboré à l'Institut de la Communication Parlée. Dans un premier temps, Guiard-Marigny (1992) a approximé, par un nombre réduit d'équations mathématiques simples, les contours des lèvres, vues de face correspondant aux 23 formes maxillo-labiales ou "visèmes" mises en évidence par Benoît et al. (1992) pour la production de la parole en français. Il a ensuite réduit, par analyse de données, le nombre des paramètres qui interviennent dans les équations permettant la génération des contours. Pour l'élaboration du modèle à trois dimensions, Adjoudani (1993), a recherché les équations des contours labiaux dans un plan perpendiculaire au plan frontal du modèle 2-D. Le modèle 3-D est commandé uniquement à partir de cinq paramètres anatomiques des lèvres, facilement mesurables sur des images vidéo du visage d'un locuteur : l'étirement interne (A), la séparation interne (B), la distance en une référence fixe sur la vue de profil et le point de contact de lèvres (C), ainsi que les protrusions de la lèvre supérieure (P1) et de la lèvre inférieure (P2).

2.2. LE MODÈLE DE VISAGE

Le modèle de visage utilisé est le modèle paramétrique de Parke (1974), qui a été implanté sur une machine Silicon Graphics et adapté à la parole par Cohen et Massaro (1990, 1993). Pour le système d'animation de ce visage, Le Goff (1993) a créé une interface de commande qui permet de contrôler les éléments faciaux intervenant dans le processus de la parole avec les cinq paramètres de notre modèle de lèvres, plus la position de la mâchoire. Il a ensuite intégré le modèle 3-D de lèvres au visage. La figure 1 montre le principe de commande de ce visage synthétique.

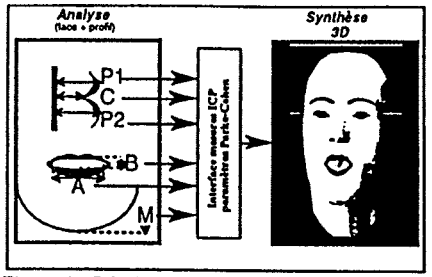


Figure 1 : Principe de l'analyse/synthèse du visage paramétrique (Le Goff, 1993).

Le Goff (1993) a quantifié l'apport d'intelligibilité à un signal de parole dégradé, dû à l'information visuelle de notre modèle. Il a montré que, quelque soit le niveau de dégradation acoustique, le visage synthétique fournit à l'auditeur presque la moitié de l'information manquante, grâce uniquement à un très faible débit d'information (6 paramètres x 25 images/sec.).

3. L'ANALYSE DES IMAGES VIDÉO

3.1. LE SYSTÈME D'ENREGISTREMENT

Pour l'enregistrement vidéo, le locuteur est placé dans une chambre sourde et filmé par deux caméras, l'une de face et l'autre de profil (Lallouache, 1991). Au début de la prise de vue, une règle est placée dans les plans de face et de profil du visage afin d'effectuer une calibration qui sera utilisée lors la phase d'analyse. Les lèvres du locuteur, ainsi qu'une zone sur le menton et des références fixes ont été préalablement maquillées en bleu pour faciliter le traitement des images vidéo : en effet un système de *chroma-key* permet de transformer, lors de la prise de vue, le bleu des lèvres en noir saturé. L'extraction des contours des lèvres peut ainsi s'effectuer plus facilement grâce à un seuillage et une binarisation de l'image. La tête du locuteur est maintenue quasiment fixe à l'aide d'un casque pour que l'image corresponde toujours au même plan. Un microphone permet l'enregistrement du signal acoustique qui est resynchronisable avec l'image à l'aide d'un système de bips, acoustiques et optiques, enregistrés simultanément. Le schéma du principe de ce système est représenté sur la figure 2.

3.2. L'EXTRACTION DES PARAMÈTRES

Le traitement du corpus enregistré peut être effectué à l'aide d'un PC muni d'une carte d'acquisition et de traitement vidéo (Matrox). Une bibliothèque de routines de traitement d'images permet d'extraire les contours des parties du visage à analyser. Les images du visage du locuteur, dont un exemple est donné sur la figure 3, sont saisies à l'aide d'un magnétoscope piloté par le micro-ordinateur. Les algorithmes développés par Lallouache (1991) permettent d'obtenir des mesures géométriques très précises mais le temps de traitement, bien que relativement court, ne permet pas d'utiliser ce poste pour une analyse des images en temps réel. Les paramètres ainsi obtenus servent à l'animation du modèle tridimensionnel de lèvres, (Guiard-Marigny, 1992; Adjoudani, 1993) ou du modèle de visage de Parke adapté aux paramètres anatomiques obtenus par l'analyse (Le Goff, 1993).

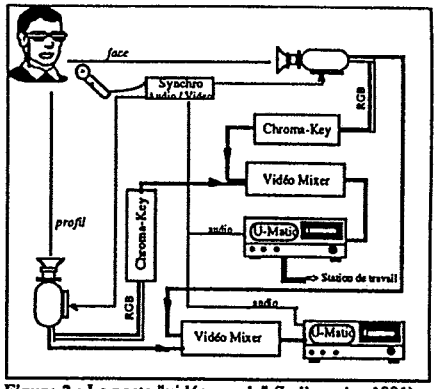


Figure 2 : Le poste "vidéo-parole" (Lallouache, 1991).

Figure 3 : Exemple d'image vidéo traitée

Le système d'analyse en temps réel est composé du même système de prise de vue et d'une station graphique (SGI-Entry). Cet ordinateur est muni d'une carte d'acquisition vidéo (*Videostarter*) permettant de saisir une image toutes les 80 ms. Le principe d'analyse des images est le même que celui utilisé par Lallouache (1991). Les algorithmes de détection des contours des lèvres et de l'extraction des paramètres ont été adaptés par Angola (1993) pour une analyse en temps réel. Le système d'analyse est composé dans un premier temps d'une phase d'initialisation. Ensuite, trois phases se succèdent : la saisie d'une image vidéo, son traitement puis l'envoi des paramètres obtenus vers la station de travail dédiée à la synthèse.

L'initialisation consiste à définir, pour un corpus donné, les fenêtres d'analyse dans lesquelles se trouvent, pendant l'enregistrement, les parties du visage dont il faut extraire les contours. La calibration effectuée à la prise de vue est alors utilisée pour obtenir des mesures en millimètres. Une fois cette phase terminée, le programme d'analyse lance un cycle de traitement. Une image est saisie et sont alors extraits les parties du visage à analyser, composées des lèvres, d'un point donnant la position du menton, ainsi que de références fixes sur les vues de face et de profil (cf. figure 4). À partir des contours ainsi traités, les mesures nécessaires pour la synthèse de notre visage, sont calculées, puis envoyées, à travers un réseau, vers une autre station graphique. Pendant ce temps, une nouvelle image est saisie, traitée et ainsi de suite. Les performances de ce programme permettent de traiter jusqu'à 12.5 images par seconde, ce qui correspond en fait à la cadence d'acquisition maximale de notre carte vidéo.

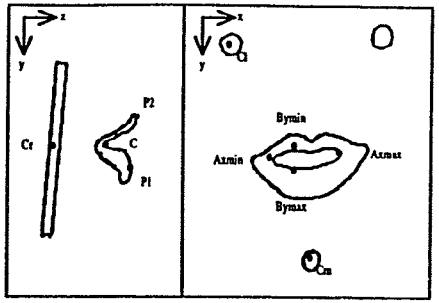


Figure 4 : Contours des objets détectés sur une image vidéo pour le calcul des paramètres à transmettre (Angola, 1993).

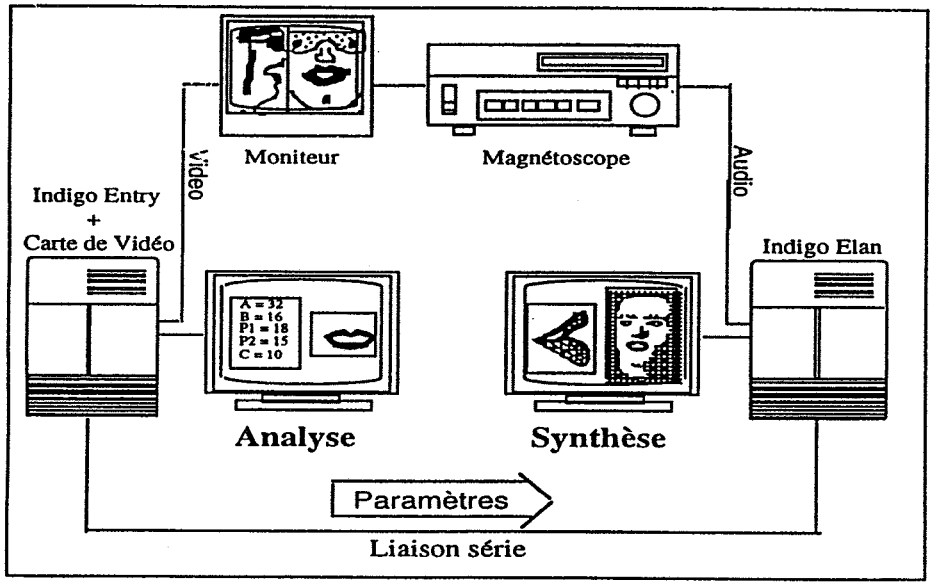


Figure 5 : Schéma de principe du système d'analyse/synthèse en temps réel d'un visage parlant (Angola, 1993).

4. L'ANALYSE-SYNTHÈSE EN TEMPS RÉEL

Le système d'analyse/synthèse en temps réel, représenté sur la figure 5, est donc implanté sur deux stations Indigo, l'une pour l'analyse, l'autre pour la synthèse. Les deux machines sont reliées entre elles par une liaison série, mais un autre type de liaison peut être envisagé. Le signal vidéo à traiter provient, soit d'un magnétoscope, soit directement du système de prise de vue. Les paramètres de commande sont transmis à une cadence de 12.5 images par seconde, par la liaison série. Le signal acoustique est le signal original, judicieusement retardé afin d'avoir une synchronisation parfaite entre le son et l'image.

La station chargée de la synthèse d'images reçoit donc un jeu de paramètres toutes les 80 ms (12.5 images par seconde). Une interpolation linéaire entre les paramètres de deux images successives permet ensuite de calculer les paramètres manquants pour une synthèse à 25 images par seconde. Cette machine permet aussi de synchroniser l'image et le son. Ainsi, elle attend de recevoir deux jeux de paramètres successifs, calcule les paramètres intermédiaires et affiche une image toutes les 40 ms. Avec ce système d'analyse/synthèse, on introduit un retard entre l'image d'origine et celle synthétisée de 200 ms (80 + 80 + 40). Il suffit donc de retarder le son correspondant de 200 ms pour synchroniser les deux processus.

5. CONCLUSION

Un prototype d'analyse/synthèse de visage parlant fonctionne en temps réel à l'ICP. Le système de synthèse utilise uniquement six paramètres anatomico-géométriques, obtenus facilement par analyse vidéo, pour l'animation du modèle de visage. Le débit nécessaire à la transmission de cette information visuelle est très faible, et autorise donc de nombreuses applications potentielles en télécommunication comme en animation automatique d'acteurs synthétiques. En outre, le principe d'analyse-synthèse ainsi mis au point permet de tester l'influence psychophysique d'un paramètre donné sur la perception bimodale de la parole., dans la mesure ou chaque paramètre mesuré sur le visage naturel peut être modifié, de façon automatique ou manuelle, avant d'être appliqué au modèle paramétrique. Dans une version "non temps-réel", ce système permet d'introduire des modifications temporelles dans la visualisation des modèles de visage, de façon à tester l'influence du débit, par exemple, sur la perception visuelle de la parole.

Une démonstration vidéo sera présentée lors de ces JEPs, et l'on pourra y voir différents "clonages" de la parole, acoustiques (en durée et en Fo) et visuels.

Références

Adjoudani, A. (1993), *Élaboration d'un modèle de lèvres 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Aizawa, K., Harashima, H., & Saito, T. (1989), *Model-Based Analysis-Synthesis Image Coding (MBASIC) System for a Person's face*, Signal Processing: Image Communication, 1, Elsevier Science Publishers, 139-152.

Angola, O. (1993), *Analyse labiométrique en temps réel par traitement d'images vidéo*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Benoît, C., Lallouache, M.T., Mohamadi, T. & Abry, C. (1992), "A set of French visemes for visual speech synthesis", in *Talking Machines: Theories, Models and Designs*, G. Bailly & C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504.

Cohen, M.M. & Massaro, D.W. (1990), "Synthesis of visible speech", *Behavioral Research Methods and Instrumentation*, 22, 260-263.

Cohen, M.M. & Massaro, D.W. (1993), "Modelling coarticulation in synthetic visual speech", Proceedings of *Computer Animation'93*, Magnenat-Thalmann & Thalmann Eds, Genève, Suisse.

Guiard-Marigny, T. (1992), *Animation en temps réel d'un modèle paramétrisé de lèvres*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat de l'Institut National Polytechnique de Grenoble, 214 pp.

Le Goff, B. (1993), *Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique, Grenoble, France.

Parke, F.I. (1974), *A parametric model for human faces*, PhD Dissertation, University of Utah, Department of Computer Sciences.

Terzopoulos, D., & Waters, K. (1991), *Techniques for Realistic Facial Modeling and Animation*, Computer Animation '91, N. Magnenat-Thalmann & D. Thalmann Eds, Springer-Verlag, 59-74

Saulnier, A., Viaud, M.-L., & Geldreich, D. (1994), *Analyse et Synthèse en temps réel du Visage pour la Télévirtualité*, Actes de la Conférence Imagina '94, Institut National de l'Audiovisuel, 173-182.

Waters, K. (1987), *A muscle model for animating three-dimensional facial expression*, Computer Graphics, 24(4), 17-24.

SYNTHESE AUDIOVISUELLE DE LA PAROLE A PARTIR DU TEXTE

Bertrand LE GOFF, Christian BENOIT

Institut de la Communication Parlée, BP 25X, 38040 Grenoble Cedex 9, France

Tél. : 76 57 76 45 – Fax : 76 82 43 35 – e-mail: legoff@icp.grenet.fr, benoit@icp.grenet.fr

ABSTRACT

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. There is evidence that synthetic faces increase the intelligibility of synthetic speech when the facial gestures and speech sounds are coherent. To reach this goal, the articulatory parameters of the facial animation have to signal the same message as the auditory speech. If a completely ambiguous message or even a contradictory message is given by the visible speech, then intelligibility would necessarily be decreased. Most of the existing parametric models of the human face have been developed in the perspective of optimising the visual rendering of facial expressions. As regards visual speech synthesis, most of the existing systems are based on a limited set of facial images occurring in the natural production of speech that are displayed one after the other, depending on the phoneme -- or rather "viseme" -- simulataneously uttered by the acoustic synthesizer. Actually, the coarticulation effects and the transition smoothing are much more naturally simulated by means of parametric models specially controlled for visual speech animation. In that perspective, a high-resolution model of the lips has been developed at the ICP. It is controlled by only five parameters. It has been implemented onto the face model originally developed by Parke and specially modified so that it is now controlled by eight articulatory parameters; the tongue and the chin motions being also rendered. Following the strategy adopted by Cohen and Massaro (1993), we have assigned target values and dominance functions to each of the eight parameters caracteristic of a French viseme. This technique allows the evolution of the parameters to be smoothed over time and gives a fair account of the coarticulation phenomenon. Our visual synthesizer has finally been synchronized with the ICP diphone-based Speech Synthesizer.

1. INTRODUCTION

Les informations visuelles apportées par le visage de l'interlocuteur contribuent de façon importante à l'amélioration de l'intelligibilité de la parole. La parole visuelle est particulièrement utile lorsque le signal auditif est dégradé (Sumby et Pollack, 1954; Erber, 1975; Massaro, 1987; Summerfield et al., 1989; Benoît et al., 1994), mais son influence ne se limite pas seulement aux situations d'audio bruité. La vidéo d'un locuteur articulant "aga" et doublée auditivement par "aba" sera perçue audiovisuellement comme "ada" (McGurk et MacDonald, 1976).

Dans le but d'améliorer l'interface homme-machine, les ordinateurs ont été dotés de voix de synthèse sans pour autant avoir un visage. La voix de synthèse n'ayant pas encore les caractéristiques de la voix humaine, elle a alors besoin d'un apport supplémentaire d'intelligibilité. Des études ont montré qu'un visage naturel pouvait reconstituer les deux-tiers de l'intelligibilité manquante dans une situation de communication où l'acoustique est dégradée (Benoît et al., 1994). Un visage synthétique en restaure la moitié (Le Goff et al., 1995). Une synthèse audiovisuelle doit donc être plus intelligible qu'une simple synthèse acoustique, à condition toutefois que le message linguistique synthétisé soit cohérent entre les deux modalités. Une telle exigence ne peut être obtenue que grâce à un modèle paramétrique de visage correctement contrôlé de façon à restituer la coarticulation naturelle de la parole. En effet, les systèmes basés sur la concaténation d'images-cibles préstockées n'ont pas l'adaptabilité nécessaire pour atteindre cet objectif. A notre connaissance, il n'existe à ce jour que trois synthétiseurs audiovisuels de la parole à partir du texte basés sur des modèles tridimensionnels de visages paramétriques : ceux d'UCSC, de l'ICP et du KTH (Benoît et al., 1995).

2. LE SYNTHETISEUR DE VISAGE PARLANT

Notre synthétiseur de visage parlant est basé sur un modèle tridimensionnel de visage à huit paramètres. Une vingtaine de "visèmes" (*visual phoneme*, Fisher, 1968) étant nécessaires pour décrire le jeu symbolique des lèvres en français (Benoît et al., 1992), un octuplet de valeurs-cible a été déterminé à partir de mesures labiométriques pour chacun de ces visèmes. Enfin, les trajectoires des paramètres "entre" ces cibles (pas forcément atteintes) sont calculées par des règles mathématiques à partir de fonctions de dominance définies par trois valeurs (intensité de l'attraction, amplitudes temporelles gauche et droite) pour un paramètre donné et un visème donné. Ces valeurs ont été calculées à partir d'observations sur un locuteur français.

Notre modèle de visage est donc animé à partir d'une chaîne de visèmes définis par leur durée. Cette chaîne de visèmes est, en particulier, facile à obtenir, par un jeu simple de réécriture, à partir d'une chaîne phonétique marquée en durée fournie par un synthétiseur acoustique.

2.1. Le modèle de visage

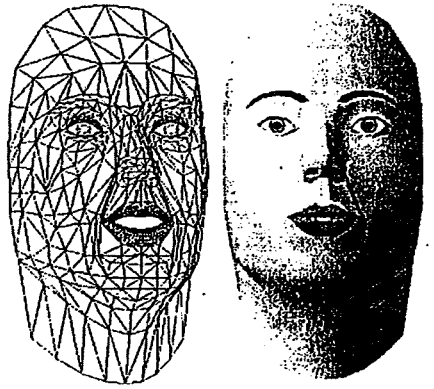


Figure 1: Modèle de visage de l'ICP. A gauche, structure fil-de-fer sans les dents ni la langue prononçant un /i/ stable. A droite, rendu par méthode Gouraud prononçant un /y/ stable.

Le modèle de visage que nous avons implanté est le résultat de la greffe du modèle tridimensionnel de lèvres élaboré à l'ICP (Guiard-Marigny et al., 1994) sur le modèle de visage originalement développé par Parke (1974) et modifié par Cohen et Massaro (1990), puis par Le Goff (1993).

Nous pilotons ce modèle à l'aide de huit paramètres, dont cinq pour les lèvres (écartement horizontal interne, séparation verticale interne, protrusion du point de contact des lèvres C, protrusion des lèvres inférieure et supérieure), un pour la position du menton (i.e. de la mâchoire), et deux pour la langue (angle de la langue avec la mâchoire, avancement de la langue).

2.2. Valeurs-cible des paramètres

Les valeurs-cible sont les valeurs qui étaient atteintes par les paramètres sans coarticulation. Ces valeurs-cible ont été mesurées sur des voyelles tenues et sur une moyenne des observations de chaque consonne dans les trois contextes i, a, y.

Nous utilisons dans cette première version dix-neuf catégories de "visèmes" : [a], [i], [y, u, ø, o, õ], [e, ɛ, ɛ̃], [ɔ], [œ, œ̃], [ã], [p, b, m], [t, d, n], [f, v], [s, z], [ʃ, ʒ], [l], [ʁ], [w], [j], forme préphonatoire (lèvres entrouvertes), forme de repos (lèvres fermées).

Ces valeurs-cible ont été mesurées à partir de mesures effectuées sur un locuteur filmé de face et de profil.

2.3. Fonction de dominance

La fonction de dominance d'un visème représente l'influence de ce visème sur ses voisins. Elle se définit à l'aide de trois coefficients (α, θ_1, θ_2) pour chacun des huit paramètres de commande : nous utilisons donc vingt-quatre coefficients. Le coefficient α_{ij} est l'intensité absolue d'attraction du visème i pour le paramètre j. Le coefficient $\theta_{1_{ij}}$ est le reflet de la croissance de la fonction : c'est la coarticulation anticipatrice. Le coefficient $\theta_{2_{ij}}$ est le reflet de la décroissance de la fonction : c'est la coarticulation rétentive. Les vingt-quatre coefficients ont été finalement obtenus par ajustements successifs de façon à approcher au plus près l'ensemble des mesures effectuées sur un locuteur.

La fonction de dominance que nous avons utilisée est l'une de celles proposées par Cohen et Massaro (1993), de la forme

$f_{ij}(t) = \alpha_{ij} . e^{-\theta_{ij}|t_0 - t|}$, où t_0 correspond au centre acoustique de la réalisation phonétique correspondante.

L'évolution temporelle du paramètre j prend alors une allure sigmoïdale entre deux cibles, dont la forme dépend des coefficients caractéristiques de chaque cible.

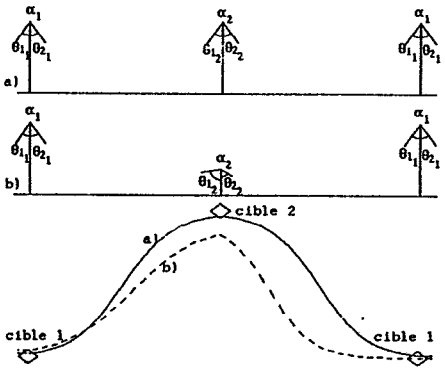


Figure 2: En haut, fonction de dominance et coefficients; en bas, trajectoire théorique d'un paramètre entre trois cibles (représentées par des losanges) Cas a) de dominances identiques. Cas b) de dominances différentes.

Sur la figure 2, nous avons représenté la trajectoire théorique d'un paramètre entre 3 valeurs-cible. Le premier cas, noté a), en trait plein, est celui de cibles de dominance identique. On voit que la cible 2 est alors atteinte. Le deuxième cas, noté b), en trait pointillé, montre la trajectoire d'un paramètre qui n'atteint pas la cible 2 parce que α_2 est trop petit. On observe donc un *"undershoot"*. Ce second exemple présente également une forte coarticulation anticipatrice (θ_{1_2} grand) et une faible coarticulation rétentive (θ_{2_2} petit) qui dissymétrise clairement la trajectoire. Ce calcul mathématique est complété par une correction de trajectoire sur le paramètre de séparation verticale afin que les fermetures des occlusives bilabiales et des fricatives labio-dentales soient effectives.

2.4. Fonctionnement du système

Les synthèses acoustique et visuelle s'effectuent en parallèle grâce au logiciel PVM (*Parallel Virtual Machine*).

La synthèse acoustique, créée par Bailly et Guerti (1991), est réalisée par concaténation d'unités stockées (polysons). L'ensemble des règles permettant la synthèse acoustique a été réalisé sous le formalisme COMPOST (Bailly et Tran, 1989). Ce COMpilateur Phonétique sur Ordinateur pour la Synthèse de Texte manipule des règles et des objets via la manipulation de listes d'atomes, d'arbres et de sous-arbres.

COMPOST analyse l'entrée (texte orthographique français) et fournit, entre autres, une chaîne de phonèmes avec leur durée, seules informations utilisées pour le moment dans notre synthétiseur de visage parlant.

Parallèlement, la synthèse visuelle effectue une conversion phonème/visème. Les trajectoires sont calculées à partir des cibles correspondant aux visèmes de la phrase. Les paramètres sont finalement stockés dans un fichier.

En sortie de synthèse, les deux fichiers visuels et acoustiques sont utilisés pour l'animation du visage synthétique avec la même technique que celle utilisée pour l'analyse-synthèse de visages réels.

3. CONCLUSION

Nous avons présenté un premier prototype de synthétiseur de la parole audiovisuelle à partir du texte. Il fonctionne en français à partir d'une entrée orthographique quelconque. Ce système a de multiples applications, notamment dans les interfaces homme-machine, l'apprentissage du français langue étrangère, ou l'apprentissage de la lecture labiale. Cependant, ce n'est qu'une première étape vers un système plus complet. Il reste à affiner les fonctions de dominance à l'aide d'un corpus plus étendu. De même, les expressions faciales, telles que clignement et mouvements d'yeux, hochement de tête, mouvement de sourcils seront implémentées et commandées par la prosodie et la syntaxe. Enfin, des tests sont en cours de réalisation afin d'évaluer l'intelligibilité de cette synthèse par rapport à un locuteur naturel.

Remerciements : Cette recherche a bénéficié du soutien du Ministère de l'Enseignement Supérieur (Bourse doctorale du MESR) et du projet européen ESPRIT-BRA N° 8579 "MIAMI".

4. BIBLIOGRAPHIE

Bailly G. et Tran A. (1989) Compost: A rule-compiler for speech synthesis, *Proceedings of Eurospeech Conference*, Paris, France, 136-139.

Bailly G. et Guerti M.(1991) Synthesis-by-rule for French, *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, n°2, 506-511.

Benoît C., Lallouache M.T., Mohamadi T. et Abry C. (1992) A set of French visemes for visual speech synthesis, *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504.

Benoît C., Mohamadi T. et Kandel S. (1994) Audio-Visual Intelligibility of French speech in noise, *Journal of Speech & Hearing Research*, n°37, 1195-1203.

Benoît C., Beskow J., Cohen M., Granstrom B., Le Goff B. et Massaro D.W. (1995) Text-to-Audio-Visual Speech Synthesis over the world, *Proceedings of the Speech Maps workshop*, C. Abry et P. Badin Eds., Grenoble, 5-6 décembre 1995.

Cohen M.M. et Massaro D.W. (1990) Synthesis of visible speech, *Behaviour Research Methods, Instruments \& Computers*, 22(2), 260-263.

Cohen M.M. et Massaro D.W. (1993) Modeling coarticulation in synthetic visual speech, *Proceedings of Computer Animation93*, Magnenat-Thalmann & Thalmann Eds, Geneve, Suisse.

Erber N.P. (1975) Auditory-visual perception of speech, *Journal of Speech & Hearing Disorders*, n°40, 481-492.

Fisher C.G. (1968) Confusions among visually perceived consonants, *Journal of Speech & Hearing Research*, n°15, 474-482.

Guiard-Marigny T., Adjoudani A. et Benoît C. (1994) A 3D model of the lips, *Proceedings of the 2nd ETRW on Speech Synthesis*, New Platz, USA.

Le Goff B. (1993) Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel, *Mémoire de D.E.A. Signal Image Parole*, Institut National Polytechnique, Grenoble, France, 106 pp.

Le Goff B., Guiard-Marigny T. et Benoît C.(1995) Read my lips ... and my jaw! How intelligible are the components of a speaker's face? *Procedings of Eurospeech'95*, Madrid, Spain.

McGurk H. et MacDonald J. (1976) Hearing Lips and Seeing Voices, *Nature*, 264, 746--748.

Massaro D.W. (1987) Speech perception by ear and eye: a paradigm for psychological inquiry., *LEA*, Hillsdale, NJ.

Parke F.I. (1974) A parametric model for human faces, *PhD Dissertation*, University of Utah, Department of Computer Sciences.

Sumby W.H. et Pollack I. (1954) Visual contribution to speech intelligibility in noise, *Journal of the Acoustical Society of America*, n°26, 212-215.

Summerfield Q., MacLeod A., McGrath M. et Brooke M. (1989) Lips, teeth, and the benefits of lipreading, in *Handbook of Research on Face Processing*, A.W. Young & H.D. Ellis Editors, Elsevier Science Publishers, 223-233.

Comment perçoit-on une femme avec un sexe masculin dans les yeux et dans les oreilles ?

C. Benoît & C. Schwippert

Institut de la Communication Parlée
Université Stendhal / INPG-ENSERG
BP 25X 38040 Grenoble Cédex, France
Tél. +33 4 76 82 43 36, Fax: +33 4 76 82 43 35, Mél: benoit@icp.inpg.fr, URL http://ophale.icp.inpg.fr

ABSTRACT

We report an experiment evaluating how discrepancies in gender between a seen face and a heard voice affect the intelligibility of acoustically degraded speech presented audio-visually to naive subjects. Scores are compared across two groups of subjects, who had been familiarized or not with one of the speakers' face and voice: Face familiarity slightly increases the benefit of lipreading.

A majority of the unfamiliar subjects identified the face alone as that of a male. A majority of them also identified the voice alone as male under highly degraded conditions. However, a majority identified a female speaker through bimodal presentation! This result challenges the models of bimodal integration based on synergetic reinforcement.

1. INTRODUCTION

Green et al. (1991) ont montré que l'effet McGurk (1976) résistait à des incohérences entre le sexe du visage et celui de la voix doublés. Cette insensibilité à la fusion des deux modalités auditive et visuelle persiste même quel que soit le degré d'incohérence, remarquée par les sujets, entre le sexe vu et le sexe entendu (Green, 1996). Walker et al. (1995) ont confirmé les résultats de Green et de ses collègues avec des sujets pour qui le visage du locuteur n'était pas familier. Par contre, elles ont montré que l'illusion McGurk était moins forte chez ceux pour qui le visage du locuteur était familier, quand il doublait une voix de sexe différent, que quand il doublait sa propre voix. Cette observation intéressante ne peut être expliquée que par une dépendance interactive entre les processus cognitifs de reconnaissance de visage et de parole.

Dans l'étude présentée ici, nous avons essayé de savoir si le "bénéfice de la lecture labiale" est significativement différent selon qu'une locutrice, dont le visage double la voix acoustiquement dégradée d'un locuteur de sexe différent, est familière ou non aux sujets. En outre, nous avons voulu évaluer jusqu'à quel point l'identification du sexe du locuteur est perturbée, quand les signaux auditifs et visuels sont ambigus.

2. LE TEST D'INTELLIGIBILITÉ

2.1. *Le corpus*

Nous avons retenu 18 logatomes de type VCVCV présentés dans une phrase porteuse de type *"C'est pas VCVCVz ?"*, où V = [a, i, y] et C = [b, ʒ, ʁ, l, v, z]. Ce corpus avait déjà été analysé dans de nombreux tests d'intelligibilité où le visage et les lèvres naturelles d'un locuteur masculin (JLS), et plusieurs modèles de visage synthétiques, étaient synchronisés au signal acoustique de JLS sous différentes conditions de bruit additif (Benoît et al., 1996).

2.2. *Les stimuli*

Une locutrice a été filmée pendant qu'elle doublait i) la voix de JLS et ii) sa propre voix, c'est-à-dire des répétitions préenregistrées de ces mêmes phrases. Une cassette vidéo présentait finalement les différents stimuli en ordre pseudo-aléatoire. Ceux-ci comprenaient des stimuli de type:

- audio seul (voix masculine [Am] ou féminine [Af]) sous deux conditions de dégradation acoustique par du bruit rose additif à S/B = -6 et -12 dB,
- visuel seul (visage féminin [Vf]) en présence de bruit rose mais sans parole,
- audiovisuel ([Vf + Am] et [Vf + Af]) sous les deux mêmes conditions de dégradation acoustique que pour [Am] et [Af] seuls.

Comme les mêmes stimuli étaient présentés visuellement cinq fois aux sujets, la locutrice portait des lunettes aveuglantes (pendant le test comme pendant la phase de familiarisation) pour éviter que des gestes non linguistiques, tels que clignement de paupières ou mouvement des yeux, fournissent des indices involontaires aux sujets pour l'identification de stimuli déjà perçus dans des conditions audiovisuelles intelligibles. Une photo du visage de la locutrice est présentée sur la figure 1.

Figure 1. Copie d'écran du visage de la locutrice.

2.3. *Les sujets*

22 étudiants en Sciences du Langage, de 18 à 28 ans ont subi le test, pour lequel ils étaient gratifiés. Aucun d'entre eux ne présentait de perte auditive ou visuelle La moitié n'avait jamais vu ni entendu les locuteurs au préalable et constituait donc le groupe NF "non familier". Les 11 autres connaissaient personnellement la locutrice et/ou ont été familiarisés avec sa voix et son visage, présenté et cadré de la même façon, à travers un enregistrement audiovisuel de 5 minutes où elle dialoguait librement sans bruit de fond.

2.4. *Le test*

Les sujets ont passé le test individuellement dans une chambre sourde en présente de l'expérimentatrice qui vérifiait que les sujets ne rataient pas les stimuli visuels. Ils étaient assis face à un moniteur couleur 27x20 cm placé à 1 m d'eux, sur lequel était placé un haut-parleur. Le niveau des stimuli acoustiques était ajusté à un niveau confortable. Les sujets étaient informés de la procédure du test par des instructions écrites. Il leur était conseillé de répondre autant que possible, même en cas de doute, et d'entourer la voyelle et la consonne perçues sur leur feuille de réponse. Les sujets du groupe F ont tous regardé la vidéo de familiarisation en préambule au test.

Une pause inter-stimuli de 7 s permettait aux sujets de de donner leur réponse. Le bruit de fond démarrait 1 s avant le début de la phrase et servait donc aussi à fixer leur attention. Les 5 premiers stimuli servaient à familiariser les sujets avec leur tâche et leurs réponses n'ont pas été analysées. En tout, la présentation des 167 stimuli durait une demi-heure.

2.5. *Résultats et discussion*

Une réponse était considérée correcte quand la voyelle et la consonne étaient identifiées correctement. Les scores globaux, moyennés sur tous les sujets de chaque groupe ont d'abord été calculés pour chacune des 9 conditions de présentation : voir figure 2.

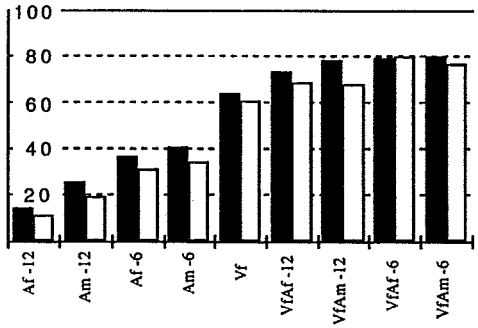


Figure 2: Scores d'identification (en %) dans chaque condition de présentation. Groupe F en noir ; NF en blanc.

Comme prévu, la parole audiovisuelle était significativement plus intelligible que l'audio seul et que le visuel seul, dans les deux groupes, à tous les niveaux de bruit (S/B = -6 ou -12 dB), et quelle que soit la voix (Am ou Af). Ce résultat confirme l'excellent qualité du doublage entre le visage et la voix.

Les différences d'intelligibilité entre les voix masculine et féminine n'étaient significatives qu'en audio seul à –12 dB, pour chacun des deux groupes ; la voix du locuteur étant plus intelligible que celle de la locutrice sous cette condition ($p<.0005$ groupe F; $p<.05$ groupe NF).

Dans l'ensemble, le groupe F a obtenu des scores plus élevés que le groupe NF ($F=17.13$, $p<.0001$). Ce résultat est vrai dans chaque condition, mais les différences n'étaient significatives que pour la voix d'homme très dégradée, dans les conditions Am-12 et VfAm-12 ($p<.005$). Comme aucun des sujets n'était familier avec la voix masculine, nous pouvons faire l'hypothèse que les différences inter-groupes ne dépendent pas de la familiarisation préalable avec le visage et la voix de la locutrice suivie par le groupe F.

Il n'y avait pas de différence entre les deux groupes dans le "bénéfice de la lecture labiale" dû à la vision du visage de la locutrice. Ce résultat va à l'encontre de nos prédictions, basées sur l'expérience de Walker et al. (1995) avec l'illusion Walker. En effet, nous nous attendions à ce que les stimuli audiovisuels construits avec la voix d'homme et le visage de femme fussent moins bien intégrés par le groupe F que par le groupe NF, le groupe F devant remarquer l'incompatibilité de sexe entre les deux modalités, alors que le groupe NF ne devait pas y être particulièrement sensible.

En fait, la dégradation acoustique était telle dans notre test (intelligibilité < 25% à –12 dB et < 40% à -6 dB quand le hasard est à 6%) que peu d'information était disponible pour la reconnaissance du locuteur. Il est même possible que le groupe F n'ait pas spécialement remarqué l'incohérence entre la voix et le visage.

Pour autant, un résultat va dans le sens de notre hypothèse. L'intelligibilité des présentations en A et AV doivent augmenter significativement quand le bruit décroît de S/B = –12 dB à –6 dB. Et c'est le cas général en effet, à part dans une seule condition : le groupe F n'a pas obtenu de scores significativement différents en audiovisuel de –12 à –6 dB quand le visage féminin doublait la voix masculine, et ce bien que l'intelligibilité soit restée loin du maximum avec seulement 80% de réponses correctes. Nous montrerons au § suivant que le sexe de la locutrice a été identifié incorrectement par une majorité des sujets du groupe NF à –12 dB en audio seul (82%), mais qu'il a été identifié correctement (à 77%) à –6 dB. Ce brusque changement catégoriel entre les deux conditions de bruit nous laisse supposer que la plupart des sujets du groupe F n'a pu remarquer l'incohérence de ces stimuli AV qu'à –6 dB, quand l'acoustique n'était pas trop dégradée, situation dans laquelle ils auraient alors moins bien intégré les deux modalités que l'autre

groupe, alors qu'ils n'ont pas eu de problèmes particuliers à plus forte dégradation, sous –12 dB. La perception d'une incompatibilité entre les deux modalités semble en effet avoir bloqué le mécanisme d'intégration bimodale...

3. L'IDENTIFICATION DU SEXE

3.1. Le protocole

À l'issue du test d'intelligibilité, les 11 sujets du groupe NF ont subi un test additionnel où ils devaient identifier le sexe du locuteur pendant une séquence de trois phrases issues du test précédent (*"C'est pas iziziz? C'est pas ububuz? C'est pas ajajaz?"*). En outre, ce test a été passé par 11 autres sujets qui ne connaissaient pas nos locuteurs. Sept d'entre eux avaient aussi subi le test d'intelligibilité, mais leurs réponses n'avaient pas été prises en compte car ils n'étaient pas francophones natifs ; les quatre autres n'ayant pas pris part au test d'intelligibilité. Les neuf conditions de présentation ont été présentées aux sujets dans l'ordre suivant : deux fois [Vf, Af-12, Am-12, Am-6, Af-6, VfAm-12, VfAf-12, VfAf-6, VfAm-6], puis Vf encore une troisième fois.

3.2. Résultats et discussion

Le pourcentage d'identification des sexes est présenté figure 3.

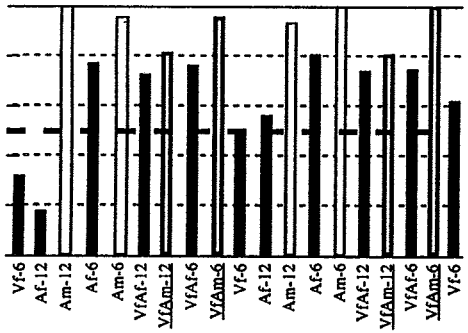


***Figure 3.** Identification du sexe en pourcentage de réponse féminine (noir) ou masculine (blanche), selon le sexe des stimuli. Les conditions incohérentes sont soulignées.*

Identification du sexe au cours du temps

Plusieurs sujets ont changé d'avis à mesure que le test progressait, puisque la deuxième série de stimuli a provoqué des réponses différentes. Beaucoup d'entre eux ont rapporté à l'expérimentatrice qu'ils ont commencé à noter une incohérence de sexe au cours de ce post-test. Ainsi, le visage seul a été identifié comme féminin par 32%, 50%, et enfin 62% des sujets de la première à la dernière présentation. Cette évolution de la perception du sexe au cours du test est telle qu'on observe même un changement catégoriel dans cette condition visuelle seule. Bien que 82% des sujets aient déjà vu ces stimuli plusieurs fois au cours du test précédent, la plupart d'entre eux ont changé d'avis quand leur attention a été focalisée vers l'identification du sexe. Ainsi, 18% des sujets ont identifié correctement le sexe de la locutrice dans la condition acoustique seule la plus dégradée [Af-12] la première fois, alors que 55% ont répondu correctement la deuxième fois.

Globalement, dans ce test d'identification du sexe, et donc a fortiori pendant le test d'intelligibilité qui le précédait, nos sujets du groupe NF ont identifié un locuteur masculin dans 66% des cas, alors que le sexe masculin n'était en réalité présent que dans 30% des stimuli.

Tous ces résultats montrent que le groupe NF a eu une forte tendance générale à considérer la parole présentée dans le test d'intelligibilité, comme produite par un locuteur masculin. Comme l'incompatibilité de sexe n'était pas évidente, il n'est pas surprenant que le groupe NF n'ait pas eu de difficulté à intégrer la parole audiovisuelle dans le test précédent.

Dominance auditive en perception du sexe

A part dans la condition VfAf-12 qui sera discutée au prochain paragraphe, l'identification audiovisuelle du sexe a largement coïncidé avec l'identification auditive. Dans deux tiers des cas où les sujets avaient perçu un sexe différent dans chaque modalité isolée, le choix bimodal suivait la décision auditive. Cette proportion s'élève à 80% dans la condition de dégradation légère (S/B = –6 dB).

Voix d'homme, visage d'homme... et mon tout est une femme !

Les réponses les plus intéressantes sont celles de la première série, avant que l'attention des sujets n'ait été dirigé vers l'incohérence de certains stimuli.

A –6 dB, 78% des sujets ont identifié correctement la voix seule de la locutrice comme féminine.

En revanche, à – 12 dB, seulement 22% d'entre eux ont identifié sa voix comme féminine.

En outre, son visage seul a été catégorisé comme féminin par 32% des sujets.

Et pourtant, ces deux dernières conditions présentées simultanément ont amené à l'identification d'une femme par 75% des sujets !

En résumé, à –12 dB, 82% M (A) + 68% M (V) = 75% F (AV) : l'intégration bimodale a amené à une catégorisation opposée à celles observées isolément dans chaque modalité !

Le visage de notre locutrice peut éventuellement être confondu avec celui d'un homme dans le cadre de cette expérience. Elle avait les cheveux courts, portait des lunettes cachant ses yeux, mais ne portait ni maquillage, ni boucles d'oreilles, ni autre "marque de féminité" particulière. Il est connu que l'absence de traits féminins tend à biaiser l'identification sexuelle vers une décision masculine.

Nous avons vu plus haut que le l'impression générale du groupe NF était qu'il y avait davantage de locuteur masculin dans le test d'intelligibilité. Ceci a certainement contribué à biaiser leur décision auditive vers une réponse masculine quand la voix de la locutrice était fortement dégradée et que les indices sexuels étaient donc très faibles. Quoiqu'il en soit, sous forte dégradation, l'intégration bimodale d'un visage perçu majoritairement comme masculin (premier stimulus) et d'une voix perçue majoritairement comme masculine (deuxième stimulus) a amené à la perception majoritaire d'une femme (sixième stimulus). Y compris au niveau individuel, où neuf sujets ont été victimes de cette surprenante illusion.

4. CONCLUSION

Le post-test qui a suivi notre test d'intelligibilité soulève un problème très intéressant en révélant une illusion originale. Il n'est pas exclus qu'elle soit (partiellement ?) due au contexte du post-test : la perception majoritairement masculine s'est produite visuellement au premier stimulus et auditivement au second. La perception féminine de l'intégration s'est produite au sixième stimulus, après que les sujets aient identifié correctement (à 78%) la voix seule, faiblement dégradée, comme féminine (sous S/B = –6 dB). Il n'est pas impossible que les sujets aient alors réalisé que deux locuteurs, de sexe différent, faisaient partie du test, et que certains aient donc soupçonné une possible incohérence dans certains stimuli. Mais cela ne suffit pas à justifier le changement de catégorie.

Notre interprétation de cette nouvelle illusion est que la parole visuelle améliore la perception de la parole à un niveau précoce, avant décodage phonétique. Non seulement la lecture labiale aide le décodage linguistique en acoustique dégradée, comme l'a encore confirmé le test d'intelligibilité précédent, mais elle met aussi en relief les caractéristiques individuelles du locuteur. Très probablement, des traits non linguistiques jouent un rôle important dans le "bénéfice moyen de 11 dB dû à la lecture labiale" suggéré par Summerfield et al. (1989) pour expliquer le bénéfice phonétique dû à la vision du visage du locuteur. Nos résultats montrent en effet qu'un visage perçu comme masculin, doublant une voix perçue comme masculine isolément, permet aux sujets de mieux percevoir les caractéristiques de cette voix, et de **l'entendre** comme si elle était produite environ 6 dB plus fort, c'est-à-dire dans une condition où cette voix a été perçue comme féminine par 16 de nos 22 sujets (la plupart des autres n'ayant donné que des réponses masculines à tous les stimuli). Cette amélioration acoustique de la voix, due à la présence du visage, a suffi ici à nos sujets pour leur faire franchir la frontière catégorielle dans la perception auditive du sexe du locuteur. La tendance à donner la réponse auditive en cas d'incohérence bimodale a fait le reste.

Cette illusion reste évidemment à être confirmée par d'autres expériences. En attendant, elle a déjà plusieurs implications. D'abord, elle soutient l'hypothèse d'une intégration précoce, pré-catégorielle, dans la perception audiovisuelle de la parole. Ensuite, elle jette un défi aux modèles d'intégration basés sur une synergie des modalités...

3. RÉFÉRENCES

[Ben96] Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996), "Which components of the face do humans and machines best speechread?", in *Speechreading by Humans & Machines*, D. Stork & M. Hennecke, Eds, Springer-Verlag, Berlin, 315-328.

[Gre91] Green, K.P., Kuhl, P.K., Meltzoff, A.M., & Stevens, E.B. (1991), "Integrating speech information across talkers, genders, and sensory modalities: Female faces and male voices in the McGurk effect", *Perception & Psychophysics*, 50(6):524-536.

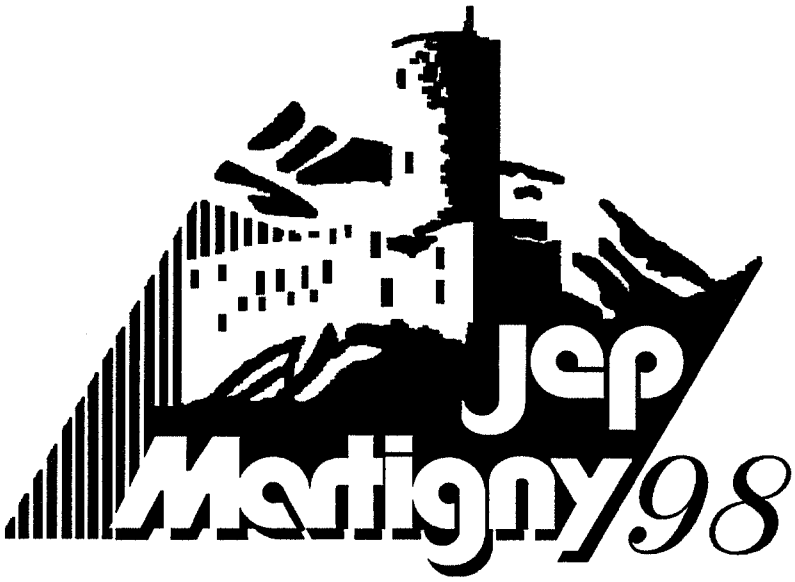
[Gre96] Green, K. (1996), "The use of auditory and visual information in phonetic perception", in *Speechreading by Humans & Machines*, D. Stork & M. Hennecke, Eds, Springer-Verlag, Berlin, 55-77.

[McG76] McGurk, H., & MacDonald, J. (1976) Hearing lips and seing voices, Nature, 264, 746-748.

[Sum89] Summerfield, Q., MacLeod A., McGrath, M., & Brooke, M. (1989), "Lips, teeth and the benefit of lipreading", in *'Handbook of research on face processing'*, A.W. Young & H.D. Ellis, Eds, Elsevier Science Publishers B.V., Amsterdam.

[Wal95] Walker, S., Bruce, V., & O'Malley, O. (1995), "Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect", *Perception & Psychophysics*, 57(8):1124-1133.

Locuteur I



Système d'élagage temps-fréquence pour l'identification du locuteur

L. Besacier, J.F. Bonastre

LIA/CERI
339, chemin des Meinajaries BP 1228 - 84911 Avignon Cedex 9 (France)
(laurent.besacier , jean-françois.bonastre)@lia.univ-avignon.fr
laurent.besacier@imt.unine.ch

ABSTRACT

This work is an attempt to refine decisions in speaker identification. A test utterance is divided into multiple time-frequency blocks on which a normalized likelihood score is calculated. Instead of averaging the block-likelihoods along the whole test utterance, some of them are rejected (pruning) and the final score is computed with a limited number of time-frequency blocks. The results obtained in the special case of time pruning, added to former results obtained in the special case of frequency pruning, lead the authors to experiment a joint time and frequency pruning approach. The optimal percentage of blocks pruned is learned on a tuning data set with the minimum identification error criterion. Validation of the time-frequency pruning process on 567 speakers leads to a significative error rate reduction on TIMIT and NTIMIT (up to 41% reduction on TIMIT) for short training and test durations. Finally, experiments in the case of simulated noise degradation have shown that this approach is a very efficient way to deal with partially corrupted speech.

1. INTRODUCTION

En reconnaissance automatique du locuteur, la modélisation mono-gaussienne du signal de parole a été largement remplacée par une modélisation multi-gaussienne (GMM) [Rey95], sensée représenter plus finement le signal. Les mixtures de gaussiennes peuvent être vues comme une *coopération* de modèles puisqu'elles reviennent à une somme pondérée de densités gaussiennes. Dans cet article, le travail présenté se rapporte plutôt à une *compétition* de modèles : différents modèles mono-gaussiens (associés à une sous-bande particulière) sont appliqués sur le signal de test et la décision est prise à partir du ou des N-meilleurs modèles.

Plus précisément, une portion de test est divisée en blocs temps-fréquence, chacun d'entre-eux correspondant à une sous-bande fréquentielle précise et à un segment temporel particulier. Pendant la phase de reconnaissance, les scores de vraisemblance calculés sur chaque bloc sont accumulés afin d'obtenir un score sur la portion de test totale. Cet article propose une stratégie d'élagage de certains scores de vraisemblance pour renforcer la décision finale.

Le but visé est de sélectionner automatiquement les parties du signal les plus aptes à identifier un locuteur. Cette approche augmente la robustesse de la décision dans le cas d'un bruit localisé sur une zone temps-fréquence précise, les blocs les moins fiables étant supprimés. Dans le cas de parole non corrompue, le procédé d'élagage permet de bâtir la décision finale sur les informations les plus pertinentes et d'ignorer les parties non informatives du signal.

Dans le *paragraphe 2*, un formalisme est proposé pour décrire le système d'élagage temps-fréquence. Son potentiel est alors mis en évidence pour le cas particulier de l'élagage temporel (*paragraphe 3*). Dans le *paragraphe 4*, nous proposons une expérience ayant pour but d'évaluer le pourcentage optimal (en terme de performances d'identification) de blocs temps-fréquence supprimés. La méthode décrite dans cet article est ensuite validée sur 567 locuteurs des bases de données TIMIT et NTIMIT (*paragraphe 5*). Enfin, dans le *paragraphe 6*, nous résumons les principaux résultats obtenus et montrons le potentiel de notre système sur des données partiellement bruitées.

2. FORMALISME

2.1 Modélisation mono-gaussienne par segments

La modélisation mono-gaussienne pour la reconnaissance automatique du locuteur est décrite plus précisément dans [Bim95] et [Gis94]. Soit $\{x_t\}_{1\le t\le M}$, une séquence de M vecteurs résultant de l'analyse *p*-dimensionnelle d'un signal de parole prononcé par le locuteur X. Ces vecteurs sont modélisés par le vecteur moyen $\bar{x}$ et la matrice de covariance X :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \quad \text{and} \quad X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \qquad (1)$$

De même, pour le signal de parole prononcé par le locuteur Y, une séquence de N vecteurs $\{y_t\}_{1\le t\le N}$ peut être extraite. En supposant que tous les vecteurs acoustiques prononcés par le locuteur X sont distribués suivant une fonction

gaussienne, la vraisemblance du vecteur y_t prononcé par le locuteur Y est :

$$G(y_t / X) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{-\frac{1}{2}(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \quad (2)$$

En admettant de plus que tous les vecteurs y_t sont des observations indépendantes, la log-vraisemblance moyenne de $\{y_i\}_{t+1\le i\le t+T}$ sur un segment de T trames, peut être écrite :

$$\overline{G_X}(y_{t+1}^{t+T}) = \frac{1}{T}\log G(y_{t+1}\ldots y_{t+T} / X) = \frac{1}{T}\sum_{i=1}^{T}\log G(y_{t+i} / X) \quad (3)$$

2.2 Modélisation multibande

Le modèle du locuteur X à 'K-sous-bandes', peut être obtenu à partir d'un modèle mono-bande initial :

$$M_X(K) = \left\{\left(X^1,\bar{x}^1\right),\ldots,\left(X^k,\bar{x}^k\right),\ldots,\left(X^K,\bar{x}^K\right)\right\} \quad (4)$$

où le locuteur X est modélisé sur la k-ième sous-bande avec la matrice de covariance X^k et le vecteur moyen $\bar{x}^k$.
X^k est un sous-bloc de la matrice de covariance X et $\bar{x}^k$ est un sous-vecteur du vecteur moyen $\bar{x}$. Ainsi, les quantités définies dans *(2)* et *(3)* peuvent être respectivement écrites sur la k-ième sous-bande :
- $G^k(y_t / X)$ vraisemblance du vecteur acoustique y_t sur la k-ième sous-bande,
- $\overline{G_X}^k(y_{t+1}^{t+T})$ log-vraisemblance moyenne du segment $\{y_i\}_{t+1\le i\le t+T}$ sur la k-ième sous-bande.

2.3 Système d'élagage temps-fréquence

Le système final combine les deux aspects *segmental* et *multibande*. Une portion de test de N trames est divisée en K sous-bandes et *n* segments de T trames chacun, tel que *N=nT (Fig. 1)*. Bien que non représenté sur la *Fig. 1*, un recouvrement entre les sous-bandes ou entre les segments est possible. Pour chaque couple (t,k) correspondant à la k-ème sous-bande du t-ième segment, un score normalisé [Gis94] $h_X^k(y_{t+1}^{t+T})$ (homogène au moins logarithme d'un rapport de vraisemblance) est calculé

$$h_X^k(y_{t+1}^{t+T}) = \max_{Z\neq X} \overline{G_Z}^k(y_{t+1}^{t+T}) - \overline{G_X}^k(y_{t+1}^{t+T}) \quad (5)$$

h est également appelé 'fonction discriminante' [Fuk90] (p.52) puisque si h<0, la vraisemblance du locuteur X est plus grande que celle de tous les autres locuteurs sur le bloc (t,k), et X est reconnu sur ce bloc. Si h>0, le locuteur reconnu sur le bloc n'est pas le locuteur X.
Les (n*K) scores normalisés sont accumulés afin d'obtenir le score de chaque modèle de référence :

$$\tilde{h}_X(y_1^N) = \underset{\substack{t\in[0,n-1]\\k\in[0,K-1]}}{ACC}\left[h_X^k(y_{tT+1}^{tT+T})\right] \quad (6)$$

ACC est la fonction d'accumulation ; notons que $h_X(y_1^N)$ est strictement équivalente à la mesure mono-gaussienne standard (bande totale) lorsque n=1 et K=1 (i.e. portion de test considérée dans sa globalité).

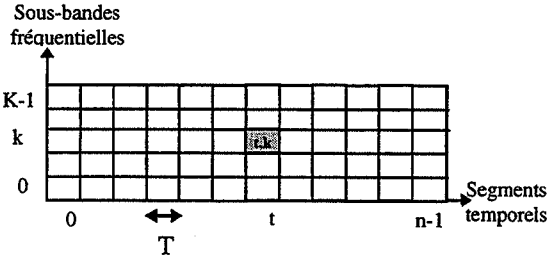


Figure 1 : Division d'une portion de test en *n* segments de K sous-bandes (*n*K* blocs au total)

L'utilisation de différents blocs temps-fréquence nous permet de supprimer ou désaccentuer des blocs correspondant à un événement anormal ou des blocs peu représentatifs du locuteur cible. La fonction d'accumulation profitant de cette segmentation est donnée dans *(7)* :

$$\tilde{h}_X(y_1^N) = \arg\min_{(p,q)}\left[\frac{1}{pq}\sum_{1}^{p}\sum_{1}^{q} h_X^k(y_{tT+1}^{tT+T})\right]_{\substack{t\in[0,n-1]\\k\in[0,K-1]}} \quad (7)$$

Dans ce cas, un score est obtenu avec la moyenne des *p*q* plus petits scores de blocs pour chaque locuteur, avec *p<n* (*n* nombre de segments dans une portion de test) et *q<K* (K nombre de sous-bandes au total). Pour une même portion de test, les blocs sélectionnés peuvent varier d'un locuteur de référence à l'autre.
Finalement, deux cas particuliers peuvent être dérivés de ce formalisme général :
- si *K=1* et *n>1*, il s'agit d'une méthode d'identification avec normalisation des scores au niveau segmental [Mar96] ; seul l'élagage temporel est alors considéré.
- si *n=1* et *K>1*, il s'agit d'une approche 'multibande' [Bes97] et seul l'élagage fréquentiel est considéré.

3. ELAGAGE TEMPOREL

3.1 Conditions expérimentales

Pour les expériences présentées ici, nous avons utilisé les bases TIMIT (parole normale) et NTIMIT (parole téléphonique). Le module d'analyse acoustique caractérise le signal toutes les 10ms par un vecteur de 24 coefficients banc de filtres. Les conditions d'analyse sont identiques à celles décrites dans [Bim95] et [Bes97]. Pour la base TIMIT, nous gardons les 24 coefficients des vecteurs spectraux. Pour NTIMIT, nous supprimons les 2 premiers coefficients et les 7 derniers qui sont en dehors de la bande téléphonique (300-3400Hz).
Le protocole d'apprentissage et de test est commun à toutes les expériences [Bes97]. Nous utilisons des durées courtes (6s d'apprentissage et 3s de test) dans le but de montrer

l'intérêt de la procédure d'élagage même lorsque peu de parole est disponible. Tous les tests sont réalisés dans le cadre d'une identification du locuteur en ensemble fermé indépendante du texte. La règle de décision est la règle du maximum de vraisemblance.

3.2 Elagage des trames

Lorsque *K=1*, le système fonctionne en mode mono-bande et seul l'élagage temporel est possible. Dans ce cadre particulier, l'influence du nombre *p* de segments conservés est étudiée lorsqu'un segment est composé d'une seule trame (*T=1*). Les résultats sont reportés dans la *Fig. 2*.

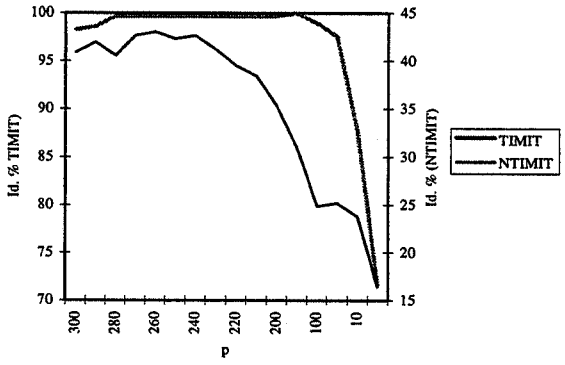


Figure 2 : Elagage temporel - 6s App./3s Test - (300-p) trames éliminées - T=1 - 63 locuteurs - 286 tests

Pour les deux bases, les résultats optimaux sont obtenus quand des trames sont supprimées : id.=100% pour *p=150* sur TIMIT et id.=43% pour *p=260* sur NTIMIT. Ceci montre que la sélection de l'information est importante puisque certaines trames de test peuvent contaminer le score final. De plus, il est intéressant de noter qu'une performance acceptable est obtenue sur TIMIT quand une seule trame est conservée (choix fait *a priori*) pour le calcul de la vraisemblance de chaque locuteur de référence (71.63% id.).

3.3 Les trames rejetées sont elles toujours les mêmes ?

Lors d'une mesure de ressemblance entre un signal de test et le modèle d'un locuteur donné, les trames conservées sont sélectionnées suivant le modèle concerné. Les scores de vraisemblance ne sont donc pas forcément tous calculés à partir des mêmes segments temporels. La *Fig. 3* montre la distribution des trames suivant leur fréquence de sélection. Ainsi, lorsque *Nselect=63* (resp. *Nselect=0*), les trames correspondantes sont conservées (resp. rejetées) par chacun des 63 modèles.

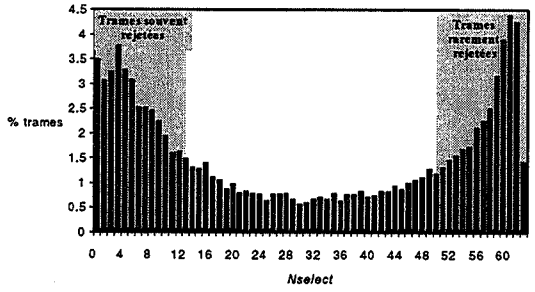


Figure 3 : Distribution des trames suivant leur fréquence de sélection - TIMIT - p=150 (la moitié des trames rejetées) - 63 locuteurs

Le profil des résultats suggère une certaine cohérence dans l'information convoyée par les trames puisque si une trame est rejetée, elle l'est par la majorité des modèles.

4. ELAGAGE TEMPS-FREQUENCE

4.1 Elagage fréquentiel

Une étude précédente [Bes97], a montré l'intérêt d'une procédure de sélection fréquentielle de l'information. Il est apparu que certaines bandes perturbaient le résultat final. De même, la quantité d'information spécifique du locuteur semble très inégalement répartie dans le spectre fréquentiel : des différences de performances allant de 25% d'identification à 5% d'identification ont pu être mesurées sur TIMIT pour des bandes constituées de 4 canaux consécutifs.

4.2 Elagage temps-fréquence

Cette expérience est réalisée avec une architecture en 24 sous-bandes de 20 canaux (architecture 24x20) pour TIMIT et une architecture en 15 sous-bandes de 11 canaux (15x11) pour NTIMIT. Les autres conditions expérimentales sont les mêmes que celles décrites dans la *section 3.1*. La taille d'un segment est T=1 (i.e. 1 segment=1 trame). Pour une durée de test de 3s (300 trames), le nombre total de blocs temps-fréquence est donc 7200 (300*24) pour TIMIT et 4500 (300*15) pour NTIMIT. L'influence, sur les performances, du nombre de blocs sélectionnés *pq* est étudiée. Les résultats sont reportés dans la *Fig. 4*.
Pour les deux bases, les meilleurs résultats sont obtenus lorsque des blocs sont supprimés : id.=100% pour *pq=3500* ou *4500* sur TIMIT et id.=41.95% pour *pq=3900* sur NTIMIT. Cependant, dans ces conditions d'expérience, il est difficile de voir le réel bénéfice d'une approche conjointe d'élagage temps-fréquence par rapport à une approche d'élagage uniquement temporel.

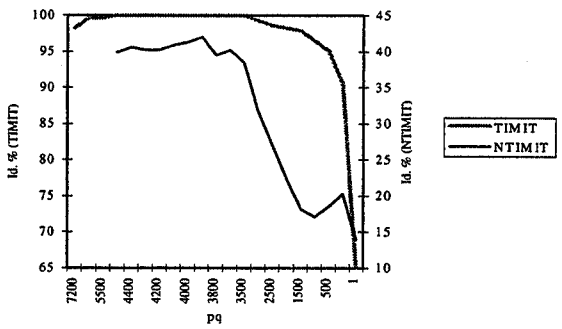


Figure 4 : Elagage temps-fréquence - 6s App./3s Test - architecture 24x20 pour TIMIT et architecture 15x11 pour NTIMIT - T=1 - 63 locuteurs - 286 tests

5. VALIDATION

Les valeurs des paramètres *p* (élagage temporel, *Section 3.2*) et *pq* (élagage temps-fréquence, *Section 4.2*) sélectionnées précédemment (à partir d'un sous ensemble composé de 63 locuteurs, de TIMIT et NTIMIT) ont été utilisées pour valider le bénéfice de la procédure d'élagage. Des tests d'identification du locuteur ont été réalisés sur un ensemble de validation distinct de l'ensemble de réglage. Ce jeu de validation est constitué par les 567 locuteurs restant de TIMIT et NTIMIT.

Les taux d'identification obtenus sont donnés dans la *Table 1*. Pour les deux bases, l'amélioration des performances est significative. La procédure d'élagage temps-fréquence permet une réduction du taux d'erreur de 41% sur TIMIT par rapport à la mesure de vraisemblance gaussienne conventionnelle. Ce bénéfice est cependant moins évident sur NTIMIT. Ceci peut s'expliquer par le plus faible nombre de bandes fréquentielles (causé par la bande passante totale réduite). De plus, l'estimation du pourcentage de blocs à rejeter reste empirique et probablement très dépendante de la base utilisée.

Table 1 : Validation de la procédure d'élagage sur TIMIT et NTIMIT (6s App./3s test - 567 locuteurs - 2639 tests)

REFERENCE		ELAGAGE TEMPOREL	ELAGAGE TEMPS-FREQUENCE
TIMIT	n=1;K=1	K=1;p=150;T=1	K=24;pq=4500;T=1
Id. %	91.66	94.20	95.14
NTIMIT	n=1;K=1	K=1;p=260;T=1	K=15;pq=3900;T=1
Id. %	15.91	18.64	17.77

6. DISCUSSION

Cet article a présenté un système d'élagage temporel et fréquentiel pour l'identification du locuteur. Les résultats obtenus ont montré que cette technique peut augmenter significativement les performances d'un système d'identification du locuteur dans certaines conditions expérimentales.

Pour montrer la robustesse de cette approche dans des conditions dégradées, un bruit distribué aléatoirement sur le domaine fréquentiel total a été ajouté aux signaux de parole de la base TIMIT. Pour chaque trame, N canaux sélectionnés aléatoirement sont dégradés pour différents rapports signal-sur-bruit. L'expérience est réalisée avec N=2 ou 3 et RSB=10dB ou 20dB.

Table 2 : Taux d'identification dans le cas de parole corrompue par un bruit distribué aléatoirement sur le domaine fréquentiel total (63 locuteurs, 286 tests).

Nombre de canaux corrompus	RSB (dB)	REFERENCE n=1;K=1	ELAGAGE T-F pq=4500;T=1
3	10	13.28	71.67
2	10	23.07	84.26
3	20	44.4	95.1
2	20	58.04	98.25

Le potentiel adaptatif de la procédure d'élagage dans de telles conditions est illustré par les résultats de la *Table 2*. Dans chaque cas, l'élagage temps-fréquence dépasse largement, en performances, l'approche conventionnelle (*reference*).

7. BIBLIOGRAPHIE

[Bes97] BESACIER, L., BONASTRE, J.F., Subband approach for automatic speaker recognition: optimal division of the frequency domain. *In Audio- and Video-based Biometric Person Authentication,* Bigün, et. al. Eds., Springer LNCS 1206, 1997.

[Bim95] BIMBOT, F., MAGRIN-CHAGNOLLEAU, I., MATHAN, L., Second-order statistical methods for text-independent speaker identification. *Speech Communication,* n°.17(1-2), August 1995.

[Fuk90] FUKUNAGA, K., *Statistical Pattern Recognition.* Second Edition, Academic Press, Inc., San Diego. 1990.

[Gis94] GISH, H., SCHMIDT, M., Text independent speaker identification. *IEEE Signal Processing Magazine*, pp 18-32, October 1994.

[Mar96] MARKOV, K., NAKAGAWA, S., Frame level likelihood normalization for text-independent speaker identification using GMMs. *In Proc. ICSLP,* pp 1764-1767, 1996.

[Rey95] REYNOLDS, D.A., Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication,* vol. 17, pp 91-108, August 1995.

Informations dynamiques et méthodes statistiques du second ordre pour l'identification du locuteur

Corinne FREDOUILLE, Jean-François BONASTRE

LIA/CERI - Agroparc - 339, chemin des Meinajaries BP1228 - 84911 Avignon Cedex 9 (France)
e-mail : (corinne.fredouille,jean-francois.bonastre)@lia.univ-avignon.fr

Abstract

In automatic speaker recognition, the use of dynamic information - stemming from the concatenation of successive frames of the speech signal - by static methods involves a computation complexity difficult to overcome if a large time window is considered. To solve this problem, we propose an original approach, whose the general idea is to split the whole frequency domain and to process dynamic information on different subbands. This method allows to reduce the information amount used and consequently the computation complexity, which makes the integration of dynamic information into the static model easier.

However, despite this improvement, the important quantity of features involved by the size of time window leads to high redundancy in models used and system performance degradation. To deal with this problem, we propose to investigate and to select the useful dynamic information part related to the time window.

Results of different experiments allowed to show the dynamic nature of information used by our identification system as well as the interest of this approach to model the dynamic characteristics of speaker illustrated by the significant performance increase observed on a majority of subbands.

1. Introduction

De nombreux travaux ont été consacrés à l'exploitation des informations dynamiques dans les systèmes de reconnaissance du locuteur. Deux approches différentes émergent des solutions proposées. La première approche consiste à extraire les dérivées (coefficients Delta et Delta-Delta) des paramètres instantanés lors de la paramétrisation [Fur81][Mag97]. Différentes études exploitant ce principe [Soo88][Ber90] ont montré que seule l'utilisation conjointe de paramètres statiques et dynamiques permet d'améliorer les performances des systèmes. La seconde approche est basée sur des méthodes dynamiques : modèles prédictifs [Mon93] [Ben95] ou modèles statiques appliqués à des données dynamiques [Nak91][Hat92]. Si ces méthodes présentent l'avantage de prendre en compte à la fois la nature statique des données et leur dynamique, elles nécessitent plus de paramètres dans les modèles ce qui amène à des problèmes de complexité de calcul.

Nous proposons ici une nouvelle méthode pour l'exploitation de paramètres dynamiques - issus de la concaténation de plusieurs trames successives de signal - par un modèle statique. Cette approche repose sur le découpage du domaine fréquentiel en sous bandes de fréquences. Un score de vraisemblance est estimé indépendamment sur chacune des sous bandes puis les résultats sont recombinés pour fournir une décision finale [Bes97] [Bes98]. Grâce à cette méthode, l'exploitation des informations dynamiques est réalisée localement au niveau de chaque sous bande fréquentielle, réduisant fortement la quantité de coefficients à traiter et permettant l'utilisation d'un nombre de trames successives plus important (empan temporel de l'ordre de 10 trames soit, dans notre cas, 100ms de signal de parole).

La section 2 présente le modèle dynamique et introduit la nécessité d'une sélection des informations au sein de chaque sous bande. Les sections 3 et 4 regroupent les protocoles expérimentaux et les résultats obtenus sur chacune des sous bandes dans le cadre d'un système d'identification automatique du locuteur. Finalement, la section 5 conclut sur la pertinence de cette méthode dynamique.

2. « Modélisation Dynamique »

2.1 Modèle statique

Le système d'identification est basé sur les méthodes statistiques du second ordre (MSSO) [Bim95][Mag97], classiquement utilisées pour modéliser les informations statiques.

A partir d'un signal de parole prononcé par un locuteur X, on extrait $\{x_t\}_{1 \le t \le M}$ une séquence de M vecteurs de dimension p, modélisés par leur vecteur moyen $\bar{x}$ et leur matrice de covariance X_0 :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \text{ et } X_0 = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T$$

De même façon, le signal de test y est modélisé par sa moyenne et sa matrice de covariance. La mesure de ressemblance utilisée pour comparer deux modèles est une variante symétrisée du maximum de vraisemblance $\mu_{G_{beta}}$ [Bim95].

2.2 Adaptation au dynamique

Dans le modèle statique précédent, chaque vecteur de dimension p issu de l'analyse acoustique est associé à une trame du signal de parole (l'ensemble des p paramètres statiques considérés est noté E_{Stat}).

Dans l'approche dynamique proposée, T spectres consécutifs sont utilisés pour représenter un empan temporel suffisant pour l'exploitation des informations dynamiques (ici, un empan temporel de 10 trames de signal successives). Par conséquent, le signal est caractérisé toutes les t milli-secondes (ici t=10ms) par un vecteur de $p*T$ coefficients dynamiques provenant de la concaténation de T vecteurs spectraux consécutifs

(fenêtre de T trames avec un recouvrement de T-1 trames avec le vecteur précédent). L'ensemble formé par ces paramètres est noté E. L'adaptation des MSSO à ce type de paramètres consiste à remplacer :

- le vecteur moyen $\bar{x}$ de dimension p par le vecteur moyen de dimension $p.T$.
- la matrice de covariance X_0 (de dimension $p*p$) par la matrice bloc-Toeplitz (de dimension $p.T*p.T$) représentant la matrice de covariance des séquences de T trames successives.

2.3 Approche multi-bandes

Même si les MSSO semblent être bien adaptées au traitement de l'ensemble E des coefficients dynamiques, la complexité de calcul engendrée par une telle quantité de données (comparée à E_{Stat}) est difficile à gérer. Le découpage du domaine fréquentiel en S sous bandes de fréquences traitées indépendamment permet de réduire d'un facteur S la quantité de données prises en compte par le modèle statique et de restreindre la complexité de calcul (si aucun recouvrement n'est appliqué). Dans ce cas, l'ensemble initial E devient :

$$E = \left\{ x / x \in E_{sb}, \forall sb \in [1,S] \right\}$$

E_{sb} étant des sous ensembles de $(p/S)*T$ coefficients dynamiques x. D'une manière similaire, des sous ensembles de paramètres statiques notés E_{sbStat} seront générés à partir de E_{Stat}.

2.4 Sélection des paramètres dynamiques

La taille de l'empan temporel souhaité impose un nombre de paramètres important, ce qui introduit des problèmes d'ingénierie non négligeables. Ce point est résolu en grande partie par le traitement individuel de chacune des sous bandes.

Cependant, la mise en évidence de l'information dynamique (présente dans les corrélations entre des coefficients ne provenant pas de la même trame) est pénalisée par la redondance de l'information statique portée par chacun des vecteurs spectraux concaténés au sein d'une fenêtre temporelle. De plus, toutes les combinaisons de coefficients ne sont pas porteuses d'information dynamique spécifique du locuteur.

Afin de résoudre ce problème, il est nécessaire de sélectionner, par sous bande de fréquences, le meilleur sous ensemble (noté E_{sbBest}) de coefficients dynamiques issus des $(p/S)*T$ paramètres de E_{sb}.

Pour cela, une variante de la méthode Knock-out [Sam75], nommée méthode de sélection Ascendante [Cha97] est utilisée. A chaque itération, cette méthode évalue tous les sous ensembles $S(n+1)$, construits en ajoutant à $S(n)$ un paramètre parmi les p coefficients restant (à la première itération $S(n) = \varnothing$ et p est égal au nombre de coefficients potentiels N). Le meilleur sous ensemble $S(n+1)$ est sélectionné et la procédure est ré-itérée ($n = n+1$ et $p = p-1$) jusqu'à ce que tous les coefficients aient été finalement ajoutés.

Le sous ensemble de coefficients dynamiques ayant obtenu les meilleures performances (suivant le critère de sélection choisi) est appelé E_{sbBest}.

Deux critères de sélection différents sont appliqués pour l'évaluation de chacun des sous ensembles potentiels :

- Taux d'identification : CritId

Ce critère correspond au rapport entre le nombre de tests d'identification réussis et le nombre total de tests effectués. Il permet de mesurer les performances des sous bandes indépendamment les unes des autres. On note $E_{sbBestId}$ les sous ensembles de coefficients dynamiques sélectionnés suivant ce critère.

- Taux de confusion : CritConf

Ce critère évalue la probabilité pour que D(x,Y) < D(x,X), avec :

D(x,Y), distance estimée entre x, un échantillon de test prononcé par un locuteur donné et Y, la référence d'un autre locuteur (mesure inter-locuteurs).

D(x,X), distance estimée entre un échantillon de test prononcé par un locuteur donné et X, la référence de ce même locuteur (mesure intra-locuteur).

Plus précisément, le critère correspond à $P(x-y<0)$, probabilité pour que y soit inférieure à x, étant données x et y deux variables aléatoires appartenant respectivement aux distributions des mesures intra et inter-locuteurs.

Dans cet article, chaque distribution est supposée gaussienne et est caractérisée par sa valeur moyenne et son écart type. Ces valeurs sont obtenues globalement et non locuteur par locuteur.

Le critère CritConf cherche à optimiser les performances en vue de la décision finale (ie après recombinaison des résultats obtenus sur chaque sous bande de fréquences).

Les sous ensembles de coefficients dynamiques sélectionnés par ce critère sont notés $E_{sbBestConf}$.

NB : les paramètres associés à la première trame de la fenêtre temporelle (premier vecteur spectral) sont automatiquement sélectionnés pour chaque sous bande.

3. Protocoles expérimentaux

3.1 Base de données et traitement du signal

Pour les différentes expériences, la base de locuteurs américains TIMIT [Fis86] est utilisée. Le signal est représenté par des vecteurs spectraux (issus d'une FFT) de 24 composantes réparties sur une échelle de Mel (ces paramètres sont issus de [Bim95]).

3.2 Protocole d'apprentissage et de test

L'ensemble des phrases est séparé en trois sous bases disjointes : une base d'apprentissage des références et deux ensembles de tests, le premier utilisé pour le développement et le second pour valider les méthodes proposées.

Deux protocoles différents sont utilisés :

- Protocole P1, développement, 135 tests de 6s provenant des 63 premiers locuteurs masculins.
- Protocole P2, test, 2639 tests de 3s (provenant des 567 autres locuteurs de TIMIT).

Il est à noter que les tests d'identification, quelque soit le protocole, sont toujours réalisés à partir de l'ensemble fermé de locuteurs constitué par les 630 locuteurs de TIMIT (la base de références contient TOUJOURS, pour P1 comme pour P2, 630 modèles).

3.3 Le système multi-bandes

Les expérimentations ont été menées sur six sous bandes constituées de 4 canaux consécutifs sans recouvrement entre les bandes (sous bande 1 : canaux 1 à 4 ; sous bande 2 : canaux 5 à 8 ; ... ; sous bande 6 : canaux 21 à 24).

3.4 Expérimentations

Trois séries d'expériences ont été réalisées :

- suivant le protocole P1 (phase de développement) pour la recherche des meilleurs sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ de paramètres dynamiques par sous bande de fréquences, en utilisant la méthode Ascendante (pour les deux critères de sélection proposés : CritId et CritConf).
- suivant le protocole P2 pour la validation des sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ obtenus précédemment par sous bande.
- suivant le protocole P2 pour vérifier la nature des informations prises en compte par les sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$: dynamique ou statique.

4. Résultats

Résultats des expérimentations de développement

Les tables 1 et 2 présentent les résultats des tests d'identification suivant le protocole P1 obtenus après sélection des sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ par la méthode Ascendante associée à CritId et CritConf. Pour comparaison, les résultats obtenus sans sélection (sur les 10 trames successives) sont fournis.

Commentaires (tables 1 et 2)

- L'apport des paramètres dynamiques des sous ensembles E_{sb}, en terme de taux d'identification, dépend fortement de la sous bande considérée : on peut remarquer une dégradation nette des performances sur les sous bandes 1-4 et 5-8 alors qu'une amélioration significative est observée pour les autres sous bandes.

	E_{sbStat}	E_{sb}	$E_{sbBestId}$
SB	**% Id.**	**% Id.**	**% Id.**
1-4	**40**	**34.8**	**48.9**
5-8	**13.3**	**8.9**	**18.5**
9-12	**5.2**	**10.4**	**11.9**
13-16	**22.2**	**26.7**	**32.6**
17-20	**40.7**	**45.9**	**54.8**
21-24	**36.3**	**40.7**	**48.9**

Tab 1 : Taux d'identification obtenus par sous bande après sélection suivant le critère CritId (Protocole P1).

	E_{sbStat}		E_{sb}		$E_{sbBestConf}$	
SB	**%Conf.**	**%Id.**	**%Conf.**	**%Id.**	**%Conf.**	**%Id.**
1-4	**23.6**	**40**	**23.6**	**34.8**	**20**	**30.3**
5-8	**13.8**	**13.3**	**15.6**	**8.9**	**12.7**	**15.6**
9-12	**23.6**	**5.2**	**22.7**	**10.4**	**20.9**	**10.4**
13-16	**10.7**	**22.2**	**10.6**	**26.7**	**9.5**	**29.6**
17-20	**8.9**	**40.7**	**7.6**	**45.9**	**7.4**	**48.1**
21-24	**7.6**	**36.3**	**7.5**	**40.7**	**6.8**	**40**

Tab 2 : Taux d'identification et taux de confusion obtenus après sélection suivant le critère CritConf (Protocole P1).

- Les taux d'identification obtenus par sous bande en utilisant les sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ sont nettement supérieurs d'une part à ceux obtenus avec les sous ensembles statiques E_{sbStat} (5,2% d'identification réussie à comparer à 11,9% et 10,4% sur la sous bande 9-12) et d'autre part à ceux obtenus avec les sous ensembles E_{sb}.
- L'objectif de CritId - optimiser les performances de chacune des sous bandes - est atteint. De même, CritConf obtient des taux de confusion nettement inférieurs à ceux des sous ensembles E_{sbStat} et E_{sb}.

Résultats des expérimentations de validation

La table 3 retrace les résultats des tests suivant le protocole P2 en utilisant les sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ obtenus lors des expérimentations précédentes. Pour comparaison, les résultats obtenus en utilisant les sous ensembles E_{sbStat} et E_{sb} sont fournis.

	E_{sbStat}	E_{sb}	$E_{sbBestId}$	$E_{sbBestConf}$
SB	**% Id.**	**% Id.**	**% Id.**	**% Id.**
1-4	**21.4**	**19**	**23.9**	**15.8**
5-8	**8.3**	**6.4**	**8.8**	**9.7**
9-12	**4.9**	**4.1**	**5.3**	**5.3**
13-16	**10.8**	**11.4**	**12.6**	**14.2**
17-20	**24.3**	**24.6**	**27.8**	**27**
21-24	**22.2**	**25**	**26.6**	**25**

Tab 3 : Taux d'identification obtenus avec P2, en utilisant les paramètres déterminés avec P1.

Commentaires (table 3)

- Les résultats sur les sous ensembles E_{sb} confirment les dégradations de performances sur certaines sous bandes dues à l'utilisation non sélective des coefficients dynamiques.
- D'autre part, on peut remarquer une amélioration significative des taux d'identification obtenus à partir des sous ensembles $E_{sbBestId}$ et $E_{sbBestConf}$ (6 à 21% d'amélioration pour CritId et de 12 à 29% pour CritConf).
- Il est intéressant de constater que CritConf obtient des taux d'identification meilleurs que CritId sur les sous bandes intermédiaires : 5-8, 9-12 et 13-16 et inversement sur les autres.
- Il peut être supposé que les informations à long terme présentes dans les basses et hautes fréquences suffisent à caractériser le locuteur, les informations dynamiques n'apportent rien de plus dans ce cas.

Nature des informations prises en compte

Afin de s'assurer de la nature **dynamique** des informations prises en compte par la méthode proposée, des tests de validation suivant P2 ont été conduits en procédant préalablement à un mélange aléatoire des trames temporelles du signal de parole. Cela doit entraîner une chute des performances si des informations de nature dynamique sont utilisées. Cette procédure a été conduite à la fois sur la phase d'apprentissage et sur la phase de test. Ce protocole s'inspire de [Mag97].

La table 4 répertorie les taux d'identification obtenus en intégrant les sous ensembles $E_{sbBestId}$ (sélectionnés avec

le critère CritId) avec et sans mélange des trames ainsi que ceux obtenus en utilisant les paramètres statiques E_{sbStat} après mélange des trames.

	E_{sbStat} avec Mélange des Trames	$E_{sbBestId}$ sans Mélange des Trames	$E_{sbBestId}$ avec Mélange des Trames
SB	% Id.	% Id.	% Id.
1-4	21.4	23.9	21.5
5-8	8.3	8.8	7.8
9-12	4.9	5.3	4.7
13-16	10.8	12.6	11.2
17-20	24.3	27.8	24.4
21-24	22.2	26.6	22.3

Tab 4 : Taux d'identification avec et sans mélange de trame (sous ensembles CritId et P2).

Commentaires (table 4)

- Comme attendu, les résultats issus des sous ensembles E_{sbStat} ne sont pas affectés par la destruction de la structure temporelle du signal de parole (voir table 3).
- En revanche, les résultats obtenus en intégrant les sous ensembles $E_{sbBestId}$ (la même tendance est observée en utilisant les sous ensembles $E_{sbBestConf}$) après destruction de la structure temporelle du signal montrent une nette dégradation des performances : une chute de 10 à 16% des taux d'identification réussie est observée sur les différentes sous bandes.
- La nature dynamique des informations prises en compte par les meilleurs sous ensembles de coefficients $E_{sbBestId}$ et $E_{sbBestConf}$ sélectionnés sur un empan de 100 ms de signal de parole est donc mise en évidence.
- Cette déduction est d'ailleurs confirmée par la similitude des performances obtenues à partir des sous ensembles E_{sbStat} et $E_{sbBestId}$ après mélange des trames montrant que la part restante d'information est, quant à elle, de nature statique.

5. Conclusion

La méthode « Dynamique » présentée ici, basée sur une approche multi-bandes, a permis de prendre en compte un empan temporel conséquent de l'ordre de 100ms de signal de parole (soit 10 trames successives) tout en conservant une complexité de calcul raisonnable.

De plus, les résultats des tests d'identification ont montré que l'exploitation d'un empan temporel de cette dimension par un modèle statique apporte une amélioration significative des performances, à condition d'effectuer une sélection des paramètres dynamiques utiles.

Finalement, les tests de validation sur la nature des informations prises en compte par les sous ensembles de coefficients optimaux $E_{sbBestId}$ et $E_{sbBestConf}$ ont démontré que cette méthode est bien adaptée à la modélisation des caractéristiques dynamiques du locuteur. En outre, l'utilisation des informations dynamiques amène à une amélioration significative des performances d'identification sur chacune des sous bandes de fréquences.

Il reste, à présent, à mesurer l'apport effectif du critère de sélection CritConf sur une amélioration globale des performances du système d'identification, en intégrant à ce dernier un module de recombinaison des différents reconnaisseurs associés à chacune des sous bandes.

6. Références

[Ben95] **BENNANI Y., GALLINARI P. (1995),** « Neural Networks for discrimination and modelization of speakers », *Speech Communication,* 17(1-2), pp 159-175.

[Ber90] **BERNASCONI C. (1990),** « On instantaneous and transitional spectral information for text-dependent speaker verification », *Speech Communication,* 9(2), pp 129-139.

[Bes97] **BESACIER L., BONASTRE J.F. (1997),** « Subband approach for automatic speaker recognition : optimal division of the frequency domain », *1st International Conference on Audio and Video Based Biometric Person Authentification (AVBPA),* Bigün et al. Eds, Springer LNCS 1206.

[Bes98] **BESACIER L., BONASTRE J.F. (1998),** « Frame pruning for speaker recognition », *Proceedings of ICASSP'98*

[Bim95] **BIMBOT F., MAGRIN-CHAGNOLLEAU I., MATHAN L. (1995),** « Second-Order Statistical Methods for text-independent speaker identification », *Speech Communication,* 17(1-2), pp 177-192.

[Cha97] **CHARLET D. (1997),** « Authentification vocale par téléphone en mode dépendant du texte », Thèse de Doctorat : Ecole Nationale Supérieure des Télécommunications, Paris.

[Fis86] **FISHER W., ZUE V., BERNSTEIN J., PALLET D. (1996),** « An acoustic-Phonetic Data Base », *JASA,* suppl. A, 81(S92).

[Fre97] **FREDOUILLE C. (1997),** « Caractérisation du locuteur au moyen d'informations dynamiques », Mémoire de DEA : Université d'Avignon (France).

[Fur81] **FURUI S. (1981),** « Cepstral analysis for automatic speaker verification », *IEEE Transactions on ASSP,* 29(2), pp 254-272.

[Hat92] **HATTORI H. (1992),** « Text-independent speaker recognition using neural networks », *Proceedings of ICASSP'92,* 2, pp 153-156.

[Mag97] **MAGRIN-CHAGNOLLEAU I. (1997),** « Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte », Thèse de Doctorat : Ecole Nationale Supérieure des Télécommunications, Paris.

[Mon93] **MONTACIE C., LE FLOCH J.L. (1993),** « Discriminant AR-Vector Models for free-text speaker verification », *Proceedings of Eurospeech'93,* 1, pp 161-164.

[Nak91] **NAKAMURA S., AKABANE T. (1991),** « A neural speaker model for speaker clustering », *Proceedings of ICASSP'91,* pp 853-856.

[Sam75] **SAMBUR M.R. (1975),** « Selection of acoustic features for speaker identification », *IEEE Transactions on ASSP,* 23(2), pp 176-182.

[Soo88] **SOONG F.K., ROSENBERG A.E. (1988),** « On the use of instantaneous and transitional spectral information in speaker recognition », *IEEE Transactions on ASSP,* 36(6), pp 871-879.

Vérification du locuteur dépendante du texte utilisant des classificateurs binaires.

Dominique Genoud, Miguel Moreira and Eddy Mayoraz

IDIAP, Dalle Molle Institute of Perceptual Artificial Intelligence
P.O. Box 592, 1920 Martigny, Switzerland
genoud@idiap.ch, miguel@idiap.ch, mayoraz@idiap.ch

ABSTRACT

This paper describes how a speaker verification task can be advantageously decomposed into a series of binary classification problems, i.e. each problem discriminating between two classes only. Each binary classifier is specific to one speaker, one anti-speaker and one word. Decision trees dealing with attributes of continuous values are used as classifiers. The set of classifiers is then pruned to eliminate the less relevant ones. Diverse pruning methods are experimented, and it is shown that when the speaker verification decision is performed with an a priori threshold, some of them give better results than a reference HMM system.

1. INTRODUCTION

La vérification du locuteur dépendante du texte est très prometteuse, car elle a conduit ces dernières années à la création d'un certain nombre d'applications dans les services bancaires et de télécommunications. La dépendance au texte est très importante dans ce type d'application car elle permet de contrôler ce que le locuteur prononce. Ce contrôle est effectué normalement par un système de reconnaissance de la parole qui segmente temporellement la phrase prononcée par le locuteur et permet ainsi la comparaison du mot reconnu avec un modèle du locuteur préalablement entraîné. Cependant, dans la plupart des systèmes, les scores obtenus pour chaque mot ou sous-mot de la phrase sont simplement sommés, sans tenir compte des connaissances *a priori* que l'on pourrait dégager pour chaque locuteur, comme par exemple le fait que la prononciation de certain mots discriminent mieux un locuteur que d'autres. Ce papier décrit comment extraire et exploiter cette information en utilisant un jeu de classificateurs binaires. Chacun des classificateurs sépare les données d'un seul mot pour un couple *client/anti-client*. Un client est un locuteur autorisé à utiliser le système de vérification et dont on veut stocker un modèle. Un anti-client est un locuteur différent de tous les clients, et dont les données serviront, dans la classification binaire, d'anti-classe aux données du client. Les sorties de chaque classificateur sont ensuite fusionnées en un seul score de différentes manières, puis le score est comparé à un seuil de décision (*a priori* ou *a posteriori*). Les classificateurs binaires ont déjà été utilisés en vérification indépendante du texte par Castellano et al. [2]. la méthode proposée est comparée à une approche HMM classique utilisant le log du rapport de vraisemblance.

2. BASE DE DONNÉES

La base de données Polycost [4] utilisée pour les expériences décrites ici est composée de 133 locuteurs enregistrés en plusieurs sessions sur lignes téléphoniques internationales . Les locuteurs sont de 13 pays différents. 104 locuteurs enregistrés sur 6 sessions ont été extraits pour le présent travail. Chaque session est composée de 4 séquences de 10 chiffres en Anglais (tous les chiffres de 0 à 9 prononcés dans un ordre différent pour chaque séquence). Les séquences de 10 chiffres sont identiques pour chaque locuteur et toutes les séquences ont été alignées temporellement chiffre par chiffre en utilisant un système automatique de reconnaissance de parole [4].

Les 104 locuteurs sont divisés en trois catégories : *CLI* contenant 82 locuteurs (48 hommes et 34 femmes), considérés comme les clients; *ANSP* contenant 12 locuteurs (6 h., 6 f.) appelés ici anti-clients; et *VALIMP* contenant les 10 locuteurs restants (5h., 5f.). *WORLD* est définit comme $ANSP \cup VALIMP$ et n'est utilisé que pour le système de référence.

Pour chaque client $s \in CLI$, l'entraînement et l'évaluation des classificateurs demande une séparation de 82 fois 6 sessions en données d'**entraînement**, de **validation** et de **test** effectuée de la façon suivante. La première session (4 séquences de 10 chiffres) du locuteur s constitue les données d'entraînement pour le client, les données d'imposture sont elles extraites des premières sessions des 12 anti-clients de *ANSP*. Le données de la 2ème session du client s sont utilisées comme **validation**; les données d'imposture pour la validation sont constituées de 4 séquences de 10 chiffres extraits aléatoirement parmi les 60 sessions des locuteurs de *VALIMP*. Les données de **test** sont composées des 4 dernières sessions du client s (16 séquences prononcées par le client) et de 81 sessions (324 séquences d'accès imposteur) chacune choisie dans les 4 sessions de chaque locuteur de *CLI* différent de s.

Pour toutes les expériences, la log énergie, 13 coefficients cepstraux LPC et leurs dérivées sont extraits du signal de parole. Un vecteur de 28 éléments utilisant une fenêtre d'analyse de 25.6ms est ainsi créé toutes les 10ms.

3. CLASSIFICATEURS

3.1. *Le système HMM de référence*

Deux types de HMM sont créés pour chaque chiffre. Tout d'abord un **modèle de monde** entraîné sur les données *WORLD* d'où sont extraites 58 répétitions de chaque chiffre prononcées par 22 différents locuteurs. Les paramètres du modèle sont estimés par entraînement classique – algorithme de Viterbi et réestimation de Baum-Welch [1]. Ensuite un **modèle de client**, utilisant le modèle de monde comme modèle d'initialisation, est réestimé sur les données du client. Tous les modèles ont la même structure HMM gauche-droite avec une gaussienne par état. Chaque modèle est composé d'un état par phonème et un état par transition entre phonèmes. Lors des tests, pour chaque chiffre prononcé par un locuteur, le logarithme du rapport de vraisemblance (LLR):

$$LLR_{\mathrm{sw}} = \log(L_{\mathrm{s}}) - \log(L_{\mathrm{w}})$$

est estimé, L_{s}, L_{w} étant les vraisemblances du modèle du client et du monde respectivement. La valeur LLR_{sw} pour chaque chiffre, pour une séquence de test donnée est sommée puis comparée à un seuil.

3.2. *Décomposition en classificateurs binaires*

Pour chaque client s de *CLI* une matrice D^s de classificateurs de taille $M \times N$ est construite (voir Figure 1). Chaque ligne de la matrice est associée à un mot (ici un chiffre, $M = 10$), et chaque colonne à un anti-client de *ANSP* ($N = 12$ dans ce cas). Pour chaque client un nombre total de 120 classificateurs est donc construit.

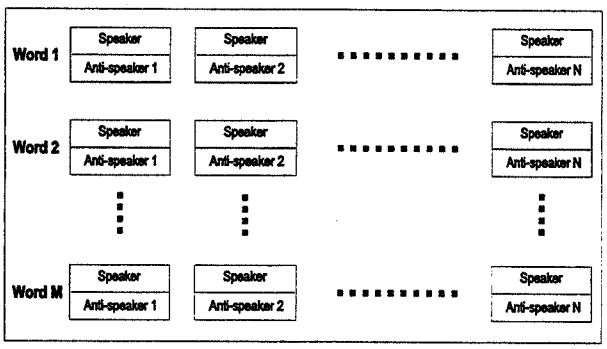


Figure 1: La Matrice D^s de classificateurs pour un client s

Lors du test d'une séquence inconnue X, chaque classificateur de D^s retourne une valeur dans $[0, 1]$ exprimant avec quelle confiance le classificateur estime que X a pu être prononcée par le client s. Les sorties des $M \cdot N$ classificateurs sont ensuite fusionnées par combinaison linéaire et comparées à un seuil t :

$$\sum_{i=1}^{M}\sum_{j=1}^{N} \omega_{ij} D_{ij}^{s}(x) \geq t.$$

Le choix des poids ω sera discuté dans la section 3.4. Le choix du seuil t sera expliqué dans la section 4. Le but de chaque classificateur est de séparer les données du client et d'un anti-client bien choisi, donc de résoudre un problème de classification à 2 classes (dit aussi classification binaire).

3.3. *Arbres de décision comme classificateurs*

Parmi les algorithmes classiques utilisés en *apprentissage automatique* pour la résolution de problèmes de classification, l'algorithme en arbre de décision C4.5 [5] , développé par Quinlan [6] a été choisi comme classificateur. C4.5 est une méthode qui permet d'utiliser des entrées avec des attributs à valeur continue. En dépit de sa relative simplicité, cet algorithme présente quelques qualités intéressantes: il a un bon pouvoir séparateur, demande très peu de temps d'entraînement et la taille d'un modèle entraîné est relativement petite (500 à 1000 bytes). Chaque classificateur dans D^s est entraîné en utilisant comme entrée les vecteurs de paramètres du client et de un des 12 anti-clients pour chacun des 10 chiffres. L'algorithme C4.5 sépare les données d'entraînement en deux classes: les données du client (classe 1) et les données de l'anti-client (classe 0). Chacun des 28 éléments du vecteur d'entrée est considéré comme un attribut continu.

Considérant les données du client et d'un anti-client, un arbre de décision est créé récursivement. Tout d'abord la racine est associée à un espace contenant toutes les données d'entraînement, puis cet espace est ensuite partitionné en sous-régions récursivement, chaque subdivision de l'espace est associée au test d'un attribut. A chaque noeud de l'arbre C4.5 on sélectionne l'attribut qui donne le meilleur gain informatif, c'est-à-dire celui qui représente le mieux la classification de sortie basée sur une **mesure d'entropie**. Le processus continue jusqu'à ce que tous les exemples de l'ensemble d'entraînement soient classés correctement, ou que tous les attributs à disposition aient été utilisés. Lorsque l'arbre est complet, chaque feuille représente une classe. C4.5 est un algorithme d'arbre de décision étendu qui peut travailler avec des données manquantes pour certains attributs. De plus l'arbre est élagué après apprentissage, permettant de remplacer une sous partie de l'arbre par une seule feuille. L'élagage est effectué lorsque le taux d'erreur escompté d'un sous-arbre est plus grande que celle d'une feuille équivalente. L'algorithme C4.5 est expliqué en détail dans [6].

Un résultat intéressant issu de l'utilisation de ce genre d'algorithme est la sélection automatique des attributs qui séparent le mieux un couple client/anti-client. En effet, le critère d'entropie permet de calculer quelle est la quantité d'information utile contenue dans chaque coefficient LPCC, ΔLPCC, énergie et Δénergie. Les coefficients sélectionnés peuvent être différents d'un classificateur à l'autre, c'est-à-dire pour chaque paire client/anti-client et pour chaque mot (chiffre).

3.4. *Fusion des décisions partielles*

La sortie de chaque classificateur fournit une décision client/imposteur. Il existe plusieurs possibilités de recombiner ces décisions partielles: la plus simple consiste à prendre la moyenne (ou la somme) des scores de tous les classificateurs de D^s. Cela peut être comparé à la somme des scores LLR dans l'approche HMM (voir sec-

tion 3.1). La différence est que chaque classificateur de D^s se concentre seulement sur la tâche de séparer le client d'un anti-client alors que dans le cas d'un système HMM utilisant un modèle de client et de monde, la séparation est effectuée entre le client et tous les autres. Parmi les $M \cdot N$ classificateurs pour un client, certains vont donner des résultats plus fiables que d'autres. Une manière d'augmenter la précision du système de vérification est de supprimer les classificateurs qui ont une réponse trop mauvaise ($\omega_{ij} = 0$) ou de pondérer ces sorties en fonction de leur qualité. Les résultats des classificateurs de D^s calculés sur les données de **validation** seront utilisées pour déterminer ω. Plusieurs techniques ont été expérimentées et sont décrites ci-dessous:

1. **Elagage à erreur zéro**
   Cette méthode sélectionne uniquement les classificateurs qui n'ont produit aucune erreur. La matrice de poids ω est une matrice binaire avec des valeurs à 1 pour les classificateurs sans erreur sur les données de validation et à 0 pour les autres.
2. **Elagage selon la distance inter-distribution**
   On utilise ici la distance entre les distributions des scores du client et des imposteurs :

   $$Dist_{\text{sp/im}} = (\mu_{\text{sp}} - \mu_{\text{im}}) - (\sigma_{\text{sp}} + \sigma_{\text{im}}) ,$$

   Où μ_{sp},σ_{sp}, μ_{im}, σ_{im} sont les moyennes et les écart-types des scores de sorties pour un classificateur binaire (i, j) du client et d'un imposteur respectivement.
   Pour un classificateur binaire, si la distance $Dist_{\text{sp/im}}$ est négative, le classificateur est éliminé ($\omega_{ij} = 0$). Sinon trois différentes approches ont été utilisées, toutes basées sur les valeurs de $Dist_{\text{sp/im}}$:
   (a) Utilisation d'un *masque binaire*, dans ce cas tous les classificateurs avec $Dist_{\text{sp/im}} > 0$ sont utilisés avec le même poids $\omega_{ij} = 1$;
   (b) Utilisation d'un *masque continu normalisé*, tel que tous les poids des classificateurs avec $Dist_{\text{sp/im}} > 0$ soient proportionnels à $Dist_{\text{sp/im}}$ et que leur somme soit égale à 1 ($\sum_i \sum_j \omega_{ij} = 1$).
   (c) *Elagage à un pourcentage d'erreur fixé.* Dans ce cas les poids utilisés sous b) sont triés en ordre décroissant, puis ils sont sommés jusqu'à ce que la somme soit supérieure à un seuil Ω. Les poids restants sont forcés à zéro ($\omega_{ij} = 0$), et finalement les poids sont re-normalisés de façon à ce que ($\sum_i \sum_j \omega_{ij} = 1$). Cette méthode d'élagage amplifie la contribution des bons classificateurs et supprime celle des plus mauvais.
3. **Elagage selon la distribution du client**
   La moyenne μ_{sp} et l'écart-type σ_{sp} de la distribution des scores du client sont calculées, les classificateurs qui ont des scores d'impostures supérieurs à $\mu_{\text{sp}} - \sigma_{\text{sp}}$ ou plus qu'un score du client inférieur à $\mu_{\text{sp}} - \sigma_{\text{sp}}$ sont supprimés. Tous les autres sont conservés ($\sum_i \sum_j \omega_{ij} = 1$).
4. **Elagage par ligne**
   Une distance identique à celle de l'élagage par estimation des distances inter-distribution puis une moyenne des classificateurs de chaque chiffre (chaque ligne de D^s) est calculée. Une pondération normalisée ($\sum_i \sum_j \omega_{ij} = 1$) dépendant des moyennes est calculée pour un chiffre. Tous les classificateurs d'un même chiffre auront la même pondération.

4. RÉSULTATS

Les résultats obtenus par le système de couplage binaire sont comparés à un algorithme HMM classique. Les classificateurs sont entraînés avec les données **d'entraînement** (comme définit dans la Sect. 2). Les tests sont effectués avec les données de **test**. Trois types de résultats sont relatés ici.

Les premiers résultats se concentrent sur les performances intrinsèques de l'algorithme de couplage binaire et du système de référence HMM, en utilisant un seuil de décision *EER (Equal Error Rate a posteriori)*. Les seconds donnent les performances des mêmes algorithmes en fonctionnement en utilisant un seuil *EER a priori* calculé sur les données de **validation**. Les résultats obtenus pour les deux types de seuils sont donnés pour les HMM, la moyenne des $M \cdot N$ classificateurs (*BP*), l'utilisation de l'élagage selon la distribution des distances avec un masque binaire (*BP-dddisc*), le même élagage avec un masque continu (*BP-ddcont*) ou un masque continu et seuil de réjection $\Omega \in 0.5, 0.9$ (*BP-ddcont-*Ω), l'élagage par ligne (*BP-digitcont*) et finalement l'élagage selon la distribution des scores du locuteur (*BP-spdisc*).

La table 1 montre les résultats obtenus avec le seuil *EER a posteriori* sur les données de **test**. Le système HMM de référence semble donner les meilleurs résultats, bien que les performances de *BP-disc* n'en soient pas très éloignées. En général, l'élagage semble détériorer les qualités intrinsèques du système.

Méthodes	FA% (6642 tests)	FR% (1312 tests)	EER%
HMM	0.24	0.26	**0.25**
BP	0.97	1.09	1.03
BP-ze	1.18	1.32	1.26
BP-spdisc	1.16	1.41	1.28
BP-dddisc	0.93	0.78	**0.85**
BP-ddcont	0.96	1.17	1.06
BP-ddcont-0.9	1.05	1.40	1.23
BP-ddcont-0.5	1.34	1.78	1.57
BP-digitcont	1.41	1.08	1.24

Table 1: Performances des différents systèmes avec un seuil de décision *a posteriori*

La table 2 montre les performances obtenues par les différents systèmes en fonctionnement avec un seuil de décision calculé à l'*EER* sur les données de **validation**. Dans ce cas l'on retrouve la dégradation de performance

classique des HMM (un taux d'erreur 20 fois plus grand ici). Cette dégradation peut être expliquée par la difficulté de séparer les distributions client/imposteur. A l'opposé quelques-uns des systèmes élagués sont beaucoup plus robustes, même si la dégradation des résultats reste non négligeable par rapport aux performances intrinsèques (de la table 1). Il est intéressant de noter que les seuls systèmes fonctionnant proche de l'EER sont les classificateurs binaires non élagués. Pour les systèmes élagués le manque d'information sur la distribution des scores de sortie pourrait expliquer les résultats non balancés.

Méthodes	FA% (6642 tests)	FR% (1312 tests)	HTER%
HMM	1.52	9.22	**5.37**
BP	11.10	0.71	5.9
BP-ze	6.20	4.60	5.40
BP-spdisc	9.39	0.62	**5.01**
BP-dddisc	8.23	0.78	**4.50**
BP-ddcont	13.01	0.93	6.98
BP-ddcont-0.9	12.89	1.25	7.07
BP-ddcont-0.5	10.12	3.20	6.66
BP-digitcont	11.80	0.62	6.21

Table 2: Performances des différents systèmes avec un seuil de décision *a priori*. HTER est l'erreur totale: (FA+FR)/2

La troisième série de résultats présentés dans la table 3 utilise un élagage binaire qui supprime la contribution de tous les classificateurs d'un chiffre (une ligne entière de la matrice D^s). Les lignes sont mises à zéro ($\omega_i = 0$) selon l'inverse de la distance calculée avec la méthode d'élagage par ligne (Section 3.4 point 4). Le nombre de lignes actives $\Delta \in 1, \ldots, 10$ est précisé (*BP-dig-*Δ). Comme pour les résultats précédents on donne les valeurs obtenues pour un seuil *a priori* (FA,FR,TER), et pour un seuil *a posteriori* (EER). Il est intéressant de constater que l'augmentation du taux d'erreur *a posteriori* n'évolue que très peu lorsque que l'on supprime les contributions des chiffres les moins discriminants. Par contre lors du fonctionnement du système avec un seuil *a priori* la dégradation des performances est beaucoup plus marquée lorsque que l'on supprime des lignes de D^s.

Méthodes	FA% (6642 tests)	FR% (1312 tests)	TER%	EER% a post.
BP	11.10	0.71	5.90	1.05
BP-dig-9	11.64	0.82	6.20	1.22
BP-dig-8	13.44	0.85	7.15	1.26
BP-dig-5	18.83	1.25	10.00	2.18
BP-dig-2	33.83	1.52	17.67	5.61
BP-dig-1	43.83	2.44	23.14	10.54

Table 3: Performances du système à suppression de chiffres avec un seuil de décision *a priori*. TER est l'erreur totale: (FA+FR)/2. EER indique le seuil *a posteriori*

5. CONCLUSIONS

Les expériences relatées dans ce papier montrent que même si intrinsèquement l'algorithme HMM de vérification du locuteur est efficace, lors de sa mise en oeuvre avec un seuil calculé *a priori* (ce qui est évidement toujours le cas lors d'applications réelles), une simple décomposition en classificateurs binaires peut être plus robuste. Une explication peut être trouvée dans la distance entre les distributions des scores client/imposteurs, cette distance semble être plus importante dans les cas des classificateurs binaires. Un effet positif des techniques d'élagage est la suppression des classificateurs qui fournissent une contribution négative à la discrimination client/anti-client. Ce n'est pas le cas de la modélisation HMM puisque la contribution de tous les locuteurs est utilisée pour le modèle de monde.

Quelques améliorations peuvent encore être apportées au système des classificateurs binaires. D'une part, différents algorithmes d'apprentissage peuvent être utilisés comme des MLP (Multi Layer Perceptrons), des arbres de décision oblique [7], ou d'autres types d'algorithmes fonctionnant bien pour la séparation en 2 classes. D'autre part, il est possible de mieux choisir le groupe d'anti-clients en essayant de décrire au mieux, dans l'espace des paramètres acoustiques, le sous-espace "non-client".

Un autre point intéressant en reconnaissance du locuteur, a été soulevé par l'utilisation des arbre de décision. En effet ceux-ci opèrent automatiquement une sélection des paramètres qui discriminent au mieux un couple client/anti-client pour un mot donné. Cela semble recouper les travaux effectué par D. Charlet et D. Jouvet [3] qui avaient déjà montré que seulement certains paramètres LPCC apportaient une contribution à la discrimination des locuteurs.

BIBLIOGRAPHIE

[1] Cambridge University Speech Group, Entropic Research Laboratories Inc., Cambridge. *HTK Hidden Markov Model Toolkit*, December 1993.

[2] P. Castelano, S. Slomka, and S. Sridharan. Telephone based speaker recognition using multiple binary classifier and gaussian mixture models. In *Proceedings ICASSP 97*, volume 2, pages 1075–1078, 1997.

[3] D. Charlet and D. Jouvet. Optimisation du paramétrage acoustique pour la vérification du locuteur. In *Acte JEP 96*, pages 399–402, 1996.

[4] D. Genoud, J. Hennebert, and H. Melin. Polycost database. In *COST250 Minutes*, 1996. POLYCOST v1.0 homepage: `http://circwww.epfl.ch/polycost/`

[5] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[6] J. Ross Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, 1993.

[7] P. E. Utgoff and C. E. Brodley. An incremental method for finding multivariate splits for decision trees. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 58–65, Los Altos, CA, 1990. Morgan Kaufmann.

Locuteur II

Reconnaissance acoustique et perceptuelle des ressemblances vocales familiales

E. Perrin, L. Collet et C. Berger-Vachon

Laboratoire Perception et Mécanismes Auditifs,
UPRESA CNRS 5020,
Pavillon U, Hôpital Ed. Herriot,
69437 LYON Cedex 03,
FRANCE
e-mail : Emmanuel.Perrin@cermep.fr

Abstract

In this work, authors focused on the effect of similar familial voice on speaker recognition. 88 couples of voice, each couple of same gender and family, were recorded twice while reading a sentence. Two speaker recognition systems were used to determine intraspeaker, intrafamily and world distances between both utterances. The first system was acoustical recognition using classical Dynamic Time Warping realignement of acoustical images composed of Linear Predictive Ceptral Coefficients. The second system was based on perceptual assessment of voices by three professionnal listeners according to a 12-criterium perceptual analysis grid. Both systems used the same Euclidean metrix to assess distances between voices. Results showed a similar behaviour of perceptual and acoustical speaker recognition systems. The acoustical system was abused by familial imposture in one case to 88. A further experiment showed that listeners where abused about one time out of four when they were requested to recognize false and true couple of familial voices.

1. Introduction

La reconnaissance du locuteur est basée sur des différences morphologiques des organes de la phonation et d'habitudes vocales entre locuteurs [1]. Ainsi le signal acoustique produit par le locuteur sera spécifique, et la reconnaissance du locuteur, acoustique (ie par des méthodes automatiques) ou perceptuelle (ie par des jurés) est possible.

Pourtant, dans le cas de liens de parenté proche (de type filiation ou fratrie) ou de vrais jumeaux, les différences morphologiques sont réduites et les habitudes vocales, très liées au contexte d'apprentissage, peuvent être très similaires.
Il n'est pas rare, dans un de ces cas, de confondre les voix d'un père et de son fils ou de deux soeurs. Dans ce cas, quels sont les mesures acoustiques qui permettent de différencier deux voix ?

Pour un système de reconnaissance automatique du locuteur (SRAL), le problème est similaire. La présentation d'un membre de la même famille ayant une voix très ressemblante s'apparente à une imposture. L'étude de l'imposture familiale peut avoir une importance dans les applications de sécurité bancaire ou de police criminelle des systèmes de reconnaissance du locuteur.
Peu d'études traitent de l'imposture familiale, et elles traitent surtout de la voix de jumeaux. Une préétude sur les voix d'un couple de jumeaux homozygote a été menée aux Laboratoires Bell [2]. Cohen et Vaich [3] se sont intéressés à la voix de 10 paires de jumeaux homozygotes. Une reconnaissance par un classificateur par minimum de distance leur a permis au mieux 95% de bonnes reconnaissances, alors que l'étude perceptuelle permettait la séparation dans 100% des cas.
Perrin et Berger-Vachon [4] ont déjà mené une préétude sur ce thème en analysant les voix de 13 couples de voix familiales du même sexe. Deux systèmes de reconnaissance du locuteur avaient été étudiés : (1) un système acoustique simple (Dynamic Time Warpping sur Coefficients Cepstraux issus de la Prédiction linéaire) et (2) une analyse perceptuelle des voix par trois jurés au moyen d'une grille d'évaluation sur 12 critères.
Les résultats avaient montré que le système acoustique et le système perceptuel avaient des comportements très semblables devant l'imposture familiale et que les SRAL pouvaient présenter une sensibilité sur la présentation d'imposteurs familiaux.

Le but de la présente étude a été, premièrement, d'étendre l'analyse de l'étude précédente [4] sur une population plus importante de locuteurs et de comparer les résultats entre le système perceptuel et le système acoustique. Deuxièmement, un jury de trois personnes compétentes a jugé du caractère familial de 88 couples de voix, dont 44 vrais couples de voix familiales et 44 couples de voix n'ayant aucun lien familial. Bien que des comparaisons de performances de reconnaissance du locuteur entre un systme acoustique et un systme perceptuel ont d'éjà été étudiées dans la littérature [5], il semble intérressant d'abor-

der ce thème dans le contexte de voix familiales.

2. Matériel et Méthodes

2.1. Base de données

Les voix de 88 paires de sujets ayant un lien de proche parenté (filiation ou fratrie) ont été considérées (Table 2.1). Pour chacun des couples de voix, les sujets étaient de même sexe.

Table 1: Couples de voix étudiés et liens de parenté

	Femmes	Hommes	Total
Nombre de sujets	54	34	88
Nombre de familles	27	17	44
Filiation	13	9	22
Fratrie	14	8	22

Deux prononciations d'une phrase du texte *La bise et le soleil* ont été enregistrées pour chacun des locuteurs. Le matériel d'enregistrement était composé d'un magnétophone REVOX B77 et d'un microphone REVOX 600 Ohms. Les enregistrements ont eu lieu dans une pièce calme. La phrase enregistrée est la suivante *Alors la bise s'est mise à souffler de toute sa force, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui, et à la fin la bise a renoncé à la lui faire ôter.*

Pour chaque sujet, les deux prononciations ont été distinguées, indépendamment de l'ordre d'enregistrement, en prononciation de référence (PR) et prononciation de présentation (PP).
Les phrases ont été acquises avec une station Unice (LIMSI-CNRS), avec une fréquence d'échantillonnage de $10kHz$ avec une quantification de 16 bits, à travers un filtre anti-repliement de fréquence de coupure $4kHz$.

2.2. Système de Reconnaissance Automatique

Le système de reconnaissance automatique est basé sur une méthode de comparaison dynamique (DTW : Dynamic Time Warping) décrite par Furui [6]. Cette méthode est dépendante du texte.

Soit $A = \{a_1, ..., a_I\}$ l'image acoustique de longueur I de la prononciation de présentation du locuteur à identifier, chaque a_i étant un vecteur de n paramètres cepstraux d'ordre n. Si $B = \{b_1, ..., b_J\}$ est l'image acoustique de la prononciation de référence et $D(i,j)$ la distance entre les vecteurs a_i et b_j, alors l'algorithme de comparaison dynamique (DTW) aligne temporellement les deux représentations A et B, et donne une distance globale $D(A,B)$.

La métrique $D(i,j)$ a été choisie comme une métrique euclidienne, comme défini dans la préétude [4] :

$$D_{i,j} = \sqrt{\sum_{k=1}^{n} (a_{i,k} - b_{j,k})^2} \qquad (1)$$

où $a_{i,k}$ et $b_{i,k}$ sont les k ièmes coefficients des images acoustiques des prononciations A et B, respectivement aux instants i et j.

Les vecteurs d'analyse étaient formés de 12 LPCC (Linear Predictive Cepstral Coefficient) issus d'une prédiction linéaire d'ordre 12. Chacun de ces vecteurs a été calculé sur une trame de 256 points de signal (25.6 ms) avec un décalage de 50% entre deux trames consécutives (12.8 ms). Chaque trame a été préaccentuée par une préemphase de 90% puis pondérée par une fenêtre de Hamming.

Chaque vecteur d'analyse a été pondéré pour éliminer les contributions de bas ordre, sensibles au bruit, et celles d'ordre élevé, pas assez discriminantes [7]. La fenêtre de pondération cepstrale a été :

$$w(i) = 1 + \frac{N sin(\frac{\pi i}{N})}{2} \qquad (2)$$

où i et N sont l'ordre et le nombre de coefficients LPCC.

2.3. Reconnaissance Perceptuelle

Les 196 prononciations ont été jugées par un jury de trois personnes compétentes (deux orthophonistes et un chercheur sur la voix) au moyen d'une grille d'analyse de 12 jugements perceptuels d'une voix :

- hauteur moyenne,
- variations de hauteur,
- intensité,
- débit,
- régularité du débit,
- nombre de pauses,
- durée des pauses,
- articulation,
- voix agréable,
- petite ou grosse voix,
- voix douce ou agressive,
- voix molle ou dynamique,
- voix monotone ou mélodieuse.

Chaque critère était codé sur une échelle variant de 0 (indice minimum) à 5 (indice maximum). Les

voix étaient jugées dans un ordre aléatoire. La distance perceptuelle $d_{a,b}$ entre la prononciation a et la prononciation b a été mesurée comme:

$$d_{a,b} = \sqrt{\sum_{i=1}^{12} (a_i - b_i)^2} \quad (3)$$

où a_i et b_i représentent le critère perceptuel i sur les voix a et b.

2.4. *Modes de présentation*

Pour les deux méthodes (acoustique et perceptuelle), chacune des prononciations de référence a été comparée à toutes les prononciations de comparaison des sujets de même sexe. Au total, 1342 comparaisons ont été effectuées : 88 de type intralocuteur (IL), 88 de type intrafamille (IF) et 1156 de type Etranger (E, seules les comparaisons d'individus de même sexe ont été effectuées).

2.5. *Analyse statistique*

La comparaison des moyennes des distances intralocuteur (IL), intrafamille (IF) et étranger (E) a été effectuée par un test t de Student de comparaison de moyennes. Seules les distances acoustiques ont été testées : les distances perceptuelles étant jugées sur une échelle discrète ne seront analysées que qualitativement. Les comparaisons IL vs IF, IL vs E et IF vs E ont été réalisées pour la méthode acoustique. Au total, plusieurs tests ont été effectués (3 tests). Pour éviter des faux positifs, la multiplicité a été prise en compte avec la correction de Bonferroni [8]. On diminue donc l'erreur de première espèce qui devient :

$$p = \frac{0.05}{\sqrt{3}} = 0.03. \quad (4)$$

3. Résultats

3.1. *Reconnaissance perceptuelle*

La table 2 présente la reconnaissance sur critères perceptuels : les distances moyennes suivent la graduation de la parenté (intralocuteur, intrafamille, étranger). Les trois groupes n'ont pas été trouvés disjoints.

Table 2: Moyenne et écart type des distances perceptuelles entre prononciations de référence et prononciations de présentation

	Intralocuteur	Intrafamille	Etranger
m	4.3	5.0	5.3
σ	2.7	2.7	3.0

3.2. *Reconnaissance Acoustique*

Les tables 3 et 4 montrent les résultats pour la reconnaissance acoustique. Le positionnement et la dispersion des groupes (Table 3 montre la même configuration que pour la reconnaissance perceptuelle. Toutefois, le groupe de distance IF semble plus proche du type Etranger que du type IL (Table 4, IF vs E, $p > 0.05$, alors que les autres comparaisons sont significatives).

Table 3: Moyenne et écart type des distances acoustiques entre prononciations de référence et prononciations de présentation

	Intralocuteur	Intrafamille	Etranger
m	0.70	0.85	0.95
σ	0.35	0.40	0.35

Bien qu'une analyse paramétrique montre des différences significatives entre les groupes, une analyse des minima et des maxima des distances par groupe a montré un chevauchement des groupes. Cela semble montrer une faible robustesse du système acoustique vis à vis de l'imposture familiale.

Table 4: Significativité des comparaisons pour la reconnaissance acoustique

	Intralocuteur	Intrafamille
Intrafamille	5.10^{-4}	-
Etranger	4.10^{-11}	> 0.05

Il ne faut pourtant pas s'arrêter à une analyse paramétrique. L'analyse des maxima et des minima des distances sur chacun des groupes montre que les groupes sont disjoints à l'exception d'un couple de voix de la même famille (sur 88). Cela montre que le sytème s'est laissé abuser une fois sur 88 par une imposture familiale. Ce résultat est important pour la robustesse à l'imposture familiale.

4. Discussion

4.1. *Analyse*

Cette étude montre un comportement semblable de la reconnaissance perceptuelle et acoustique de voix familiales. Le système peceptuel montre une séparation des groupes avec un certain chevauchement, alors que le système acoustique, montre un faible chevauchement des distances en fonction du type de comparaison. Ceci permet de penser à une plus forte robustesse du système acoustique vis à vis d'impostures familiale par rapport au système perceptuel. Ces résultats diffèrent de ceux de Cohen et

Vach [3], qui trouvaient, sur un échantillon de sujets plus réduit, une meilleure performance du coté perceptuel dans le cas de voix de jumeaux homozygotes.

4.2. Expérience complémentaire

On doit remarquer que dans notre cas, les jurés n'ont pas eu à décider de la ressemblance de deux voix, c'est à dire, qu'il n'y a pas eu de prise de décision. Pour essayer d'analyser cet effet de prise de décision, une expérience complémentaire a été menée sur la reconnaissance de vrais et de faux couples de voix familiales.

Dans cette expérience complémentaire, nous avons posé la question de la reconnaissance de deux voix au un jury : ces deux voix appartiennent-elles á la même famille ? Pour ceci, 88 couples de voix ont été sélectionnés : 44 couples de voix familiales (groupe 1) et 44 couples de voix de sujets ne présentant aucun lien de parenté (groupe 2). Chaque couple de voix était composé de sujets du même sexe et les deux groupes de voix (vrai et faux couples) ont été appariés en terme de sexe et de différence d'âge entre sujets d'un même couple. Par exemple, un couple de voix familiales (Groupe 1) représenté par une mère et sa fille, correspondait dans le groupe 2, deux femmes de même différence d'âge n'ayant aucun lien de parenté.

La table 5 montre les scores de bonne reconnaissance exprimés en pourcents. Un score de reconnaissance de 50% représente un choix au hasard. On remarque qu'en moyenne, 71% des reconnaissances sont justes et que la reconnaissance de voix familiales (76%) est plus facile que la distinction de voix étrangères (66%).

Table 5: Pourcentages de bonnes reconnaissances. VC : vrais couples, FC : faux couples.

	VC+FC	VC	FC
Bonnes Reco.	71%	76%	66%

5. Conclusion

Cette étude a montré des similarités acoustiques et perceptuelles des voix familiales. Le système de reconnaissance acoustique ne s'est laissé abuser qu'une fois sur la présentation de 88 couples de voix familiales. Le système perceptuel a montré des performances moins bonnes. Une expérience complémentaire de reconnaissance perceptuelle de vrais et de faux couples de voix familiales par un jury a montré une erreur sur quatre dans la reconnaissance.

6. Remerciements

Les auteurs remercient Mlles Arband et Magnon pour leur contribution à ce travail.

Bibliographie

[1] S. Furui, "An overview of speaker recognition technology," in ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Avril 1994, pp. 1–9.

[2] A.E. Rosenberg, "Speaker Recognition - Identifying People by their Voices," Proc. IEEE, vol. 64, pp. 1651–1663, 1985.

[3] A. Cohen and T. Vaich, "On the identification of twins by their voice," in ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Avril 1994, pp. 213–216.

[4] E. Perrin and C. Berger-Vachon, "Reconnaissances de voix familiales," in XXe JEP, 1996, pp. 403–408.

[5] M.M. Homayounpour, G Chollet, J.-P. Goldman and J. Vaissière, "Performance comparison of Machine and Human Speaker Verification," in EUROSPEECH 93, 1993, pp. 2295–2298.

[6] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamical features," IEEE Trans. Acoust. Speech Signal Proc., vol. 29, pp. 342–350, 1981.

[7] B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the use of band-pass filtering in speech recognition," IEEE Trans. Acoust. Speech Signal Proc., vol. 35, pp. 947–954, 1987.

[8] P. Dagnelie, Théorie et méthodes statistiques, vol. I, Presses Agronomiques de Gembloux, 1975.

Méthode d'estimation du seuil de décision en vérification du locuteur

Jean-Benoît PIERROT, Frédéric BIMBOT

ENST - Dépt Signal / CNRS - URA 820
46, rue Barrault - 75634 Paris cedex 13, FRANCE

pierrot@sig.enst.fr - bimbot@irisa.fr

ABSTRACT

We present an approach for estimating decision thresholds in speaker verification, based on client enrollment data only. Our preliminary results tend to show that it is possible to correct for part of the bias between the theoretical and optimal threshold without additional speech data. All experiments have been carried out following the CAVE test protocol on SESP.

1. INTRODUCTION

Une des difficultés en vérification du locuteur consiste à déterminer un seuil de décision qui permette de faire un choix entre les deux alternatives suivantes : l'identité de l'utilisateur correspond à l'identité du locuteur proclamé ou bien elle n'y correspond pas. Dans des applications pratiques, ce seuil de décision doit en outre être estimé *a priori*. Cette estimation *a priori* s'oppose, dans son principe, à une estimation *a posteriori* qui est généralement effectuée lors d'évaluation de procédés et en connaissant les identités réelles des locuteurs.

Cet article présente les hypothèses et les étapes successives de modélisation qui mènent à l'estimation de ce seuil: la modélisation de la distribution des observations émises par un locuteur (le modèle de locuteur) puis la modélisation de la réponse du procédé de vérification.

Des méthodes ont été proposées [Bim97] [Pie98] mais celles-ci nécessitent des données additionnelles à celles ayant servies à l'initialisation des paramètres du modèle de locuteur. Les travaux présentés ici visent à s'affranchir de ces données en utilisant le plus possible les données d'apprentissage du client.

Le cadre pratique de cet article est une expérimentation utilisant un modèle de locuteur basé sur les Modèles de Markov Cachés et le protocole de test est celui élaboré lors du projet européen CAVE [Bim96].

2. ASPECTS THÉORIQUES

2.1. *Notations*

On appellera X l'identité d'un locuteur, $\bar{X}$ correspondant au reste de la population. On notera par Y la variable muette d'identité de locuteur (soit X soit $\bar{X}$). La décision à la sortie du système de vérification sera notée $\tilde{Y}$.

2.2. *Rappel de théorie de la décision*

En vérification du locuteur (et plus généralement en théorie de la décision bayésienne [Sch90]), le critère que l'on cherche à minimiser est la fonction de coût total (1) (ou risque moyen de décision (DCF[1])):

$$R_{dcf} = C[\tilde{X} \mid \bar{X}].p_{\bar{X}}.P_{\bar{X}}(\tilde{X}) + C[\tilde{\bar{X}} \mid X].p_X.P_X(\tilde{\bar{X}}) \quad (1)$$

où $C[\tilde{Y} \mid Y]$ est le coût de la prise de décision $\tilde{Y}$ sachant que l'identité du locuteur est Y, p_Y est la fréquence d'occurence du locuteur Y et $P_Y(\tilde{Y})$ la probabilité de la décision $\tilde{Y}$ sous l'hypothèse que l'identité réelle est Y.

On démontre [Sch90] que la minimisation de C revient à comparer à un seuil le rapport des distributions des observations d'une part sous l'hypothèse d'identité du locuteur (X) et d'autre part sous l'hypothèse du reste du monde ($\bar{X}$) (2):

$$l(O) = \frac{F_X(O)}{F_{\bar{X}}(O)} \begin{matrix} \text{accepté} \\ > \\ < \\ \text{rejeté} \end{matrix} \frac{C[\tilde{X} \mid \bar{X}].p_{\bar{X}}}{C[\tilde{\bar{X}} \mid X].p_X} = \theta \quad (2)$$

où O est l'observation (la séquence de trames acoustiques) et $F_Y(O)$ sa distribution sous l'hypothèse d'identité Y.

2.3. *Modèle de distribution des observations*

Introduction: Les distributions des observations sont définies sous les deux hypothèses d'identité de locuteur. Lorsque ces distributions sont les représentations exactes des répartitions des observations, le seuil de décision théorique θ est optimum. En pratique elles sont modélisées imparfaitement par des fonctions de vraisemblance ($\hat{F}$) qui génèrent un biais par rapport à la théorie:

$$F \underset{\text{modélisation}}{\Rightarrow} \hat{F} \quad (3)$$

Le rapport de l'équation (2) s'écrit dans sa version modélisée comme:

[1]Detection Cost Function

$$\hat{l}(O) = \frac{\hat{F}_X(O)}{\hat{F}_{\bar{X}}(O)} \quad (4)$$

Le Modèle de Markov Caché: Pour les expériences, on a choisi d'utiliser un modèle de locuteur de type Modèle de Markov Caché (ou HMM). Cette modélisation partitionne l'espace acoustique suivant des critères stochastiques. Chaque portion de ce partitionnement est représentée par un état dont l'estimation des paramètres dépend de l'algorithme d'initialisation des chaînes et des données d'apprentissage (dont les caractéristiques sont "représentatives" du locuteur). Lors d'une tentative d'accès, un algorithme de mesure attribue à l'observation la vraisemblance d'avoir été émise par le modèle de locuteur. Particulièrement utilisé pour les Modèles de Markov Cachés, l'algorithme de Viterbi approxime cette valeur par la vraisemblance du meilleur chemin dans le modèle [You97]. Dans ce cas, la vraisemblance se décompose comme le produit des vraisemblances des observations $\hat{F}_Y(o_t \mid q_t)$ conditionnellement aux états parcourus et des probabilités de transition $\hat{F}_Y(q_t \mid q_{t-1})$ entre ces états:

$$\begin{aligned} \hat{F}_Y(O \mid Q) &= \prod_{t=1}^{T} \hat{F}_Y(o_t \mid q_t)\hat{F}_Y(q_t \mid q_{t-1}) \\ &\text{avec} \quad \hat{F}_Y(q_1 \mid q_0) = \hat{F}_Y(q_1) \end{aligned} \quad (5)$$

Dans cette équation, T est le nombre total de trames et q_t l'état attribué à l'échantillon de l'instant t par l'algorithme de Viterbi. En rassemblant les observations de l'équation (5) selon leur attribution aux états du Modèle de Markov Caché, où q est l'état étudié, o_{t_q} l'observation attribué à cet état, T_q son nombre de trames associées à l'état q et Q le nombre total d'états du modèle, on obtient:

$$\begin{aligned} \hat{F}_Y(O \mid Q) &= \prod_{q=1}^{Q} \prod_{t_q=1}^{T_q} \hat{F}_Y(o_{t_q} \mid q) \prod_{t=1}^{T} \hat{F}_Y(q_t \mid q_{t-1}) \\ &\text{avec} \quad \sum_{q=1}^{Q} T_q = T \end{aligned} \quad (6)$$

Dans le contexte de nos expériences, nous utilisons des fonctions mono-gaussiennes pour chacun des états. Le logarithme du rapport de vraisemblance s'écrit alors:

$$\begin{aligned} \hat{ll}(O) &= \log(\hat{l}(O)) \quad (7) \\ &= \frac{1}{2} \sum_{(q;\bar{q})=(1;1)}^{(Q;\bar{Q})} \sum_{t_{q\bar{q}}=1}^{T_{q\bar{q}}} \varphi(q, \bar{q}, o_{t_{q\bar{q}}}) + K \end{aligned}$$

avec

$$\begin{aligned} \varphi(q, \bar{q}, o_{t_{q\bar{q}}}) &= (o_{t_{q\bar{q}}} - \mu_{\bar{q}})^{+} \Gamma_{\bar{q}}^{-1} (o_{t_{q\bar{q}}} - \mu_{\bar{q}}) \\ &- (o_{t_{q\bar{q}}} - \mu_q)^{+} \Gamma_q^{-1} (o_{t_{q\bar{q}}} - \mu_q) \end{aligned}$$

On utilise ici la même notation que pour l'équation (6) mais en notant les attributions conjointes de l'observation $o_{t_{q\bar{q}}}$ aux états du modèle de locuteur (q) et de non-locuteur ($\bar{q}$). Le vecteur μ_q (resp. $\mu_{\bar{q}}$) est le vecteur de moyenne estimé du modèle de locuteur (resp. de non-locuteur) de l'état q et Γ_q sa matrice de covariance (resp $\bar{q}$ et $\Gamma_{\bar{q}}$ pour le non-locuteur). La constante K comprend l'ensemble des probabilités de transitions entre états et les logarithmes des rapports des déterminants des matrices de variance.

2.4. *Modèle de distribution du rapport des fonctions de vraisemblance*

Introduction: Au moment d'un test de vérification, un ensemble de trames d'observations peut être vu comme un échantillon issu d'un tirage aléatoire. Le rapport de vraisemblance est dépendant de cet échantillon. Ce rapport est donc lui-même une variable aléatoire scalaire qui suit une distribution f résultant de la transformation par la fonction de vraisemblance ($\hat{F}_Y$) de celle des observations (F_Y):

$$\begin{aligned} \hat{ll}(O) = \log \hat{l}(O) &= \log \frac{\hat{F}_X(O)}{\hat{F}_{\bar{X}}(O)} \quad (8) \\ &\rightarrow \quad f(F_X, \hat{F}_X, F_{\bar{X}}, \hat{F}_{\bar{X}}, O) \\ &\rightarrow \quad f(F_X, \hat{F}_X, F_{\bar{X}}, \hat{F}_{\bar{X}}, \hat{ll}) \end{aligned}$$

Le critère de décision (2) s'applique maintenant sur le logarithme du rapport de vraisemblance $\hat{ll}$ et $f_Y(\hat{ll})$ est sa distribution conditionnellement à Y:

$$\frac{f_X(\hat{ll})}{f_{\bar{X}}(\hat{ll})} \begin{matrix} \text{accepté} \\ > \\ < \\ \text{rejeté} \end{matrix} \theta \quad (9)$$

où θ est identique au seuil de l'équation (2). La mesure de log-vraisemblance sur les observations est réalisée après l'attribution aux états du HMM par l'algorithme de Viterbi. Il s'agit donc d'une mesure dépendante de ce tirage d'états. L'équation (10) tient compte de cet aspect composite des hypothèses:

$$\frac{f_X(\hat{ll} \mid Q\bar{Q})}{f_{\bar{X}}(\hat{ll} \mid Q\bar{Q})} \begin{matrix} \text{accepté} \\ > \\ < \\ \text{rejeté} \end{matrix} \theta . \frac{P_{\bar{X}}(Q\bar{Q})}{P_X(Q\bar{Q})} = \theta' \quad (10)$$

où $P_Y(Q\bar{Q})$ est la probabilité d'obtenir les suites d'états particulières ($Q\bar{Q}$) sous l'hypothèse que le locuteur est Y. On va aussi supposer l'égalité des probabilités d'obtention des suites d'états ($Q\bar{Q}$) indépendamment de l'hypothèse de locuteur. Ce qui donne $P_{\bar{X}}(Q\bar{Q}) = P_X(Q\bar{Q})$.

La résolution de l'équation (10) revient à déterminer [Fuk72] la valeur du seuil optimum (α) à appliquer sur $\hat{ll}$ pour obtenir les conditions de la minimisation du coût:

$$\frac{f_X(\hat{ll} = \alpha \mid Q\bar{Q})}{f_{\bar{X}}(\hat{ll} = \alpha \mid Q\bar{Q})} = \theta' \qquad (11)$$

Comme on l'a fait lors de la modélisation des distributions des observations on va chercher maintenant à définir un modèle de distribution $\hat{f}_Y(\hat{ll})$ du rapport de vraisemblance. Cette nouvelle étape de modélisation a l'avantage de se ramener à l'étude de la distribution d'une variable aléatoire scalaire plutôt que vectorielle. Certains auteurs ont choisi de fixer des hypothèses particulières sur cette distribution (gaussianité pour [Bim97]) et d'autres estiment leurs moyennes et variances [Pie98] puis déterminent le seuil de décision. Ces méthodes ont besoin de données de développement générées par des locuteurs différents des premiers mais possédant les mêmes caractéristiques statistiques. Une autre solution est l'utilisation d'un ensemble de données supplémentaires sur les locuteurs. Cette alternative nécessite une participation plus longue de la part des locuteurs lors de la phase d'apprentissage.

Dans tous ces cas, l'estimation des paramètres de distribution de la variable scalaire de vraisemblance exige un ensemble complémentaire de données. Dans cet article on tente de s'affranchir de ces données supplémentaires et on utilise les informations portées par les paramètres estimés du modèle de locuteur ($\hat{F}_Y$) pour déterminer analytiquement les paramètres de $\hat{f}_Y$. Bien entendu ce calcul analytique n'est réalisable que dans le cas de formes simples du modèle de distribution des observations. Dans des cas plus complexes, il est néanmoins possible de générer des observations artificielles à partir du modèle de locuteur puis de déterminer la distribution de la réponse du système de vérification avec ces tirages.

Le modèle gaussien: Pour nos expériences, le modèle choisi pour $\hat{f}_Y$ est gaussien. On suppose que conditionnellement à un couple d'états $(q, \bar{q})$, la somme de variables aléatoires converge vers une loi gaussienne [Sap90]. De plus la somme de ces variables gaussiennes est aussi supposée gaussienne. En se plaçant sous l'hypothèse Y, on a $\hat{ll}(O) \rightarrow \mathcal{N}_Y(m_Y; \sigma_Y)$. Grâce à cette hypothèse de gaussianité et la relation (11) on peut déterminer analytiquement la valeur de α.

Pour estimer les moyennes et les variances conditionnellement aux hypothèses, on utilise les paramètres des modèles de locuteur estimés lors de leur initialisation. Pour cela on suppose que les observations sont émises par le modèle de locuteur lorsque nous nous trouvons sous cette hypothèse et réciproquement pour le non-locuteur, ce qui donne pour la moyenne:

$$\begin{aligned} m_Y &= E_Y[\hat{ll}(O)] \\ &= \sum_{(q;\bar{q})=(1;1)}^{Q\bar{Q}} \frac{T_{q\bar{q}}}{T} E_Y[\varphi(q, \bar{q}, o_{t_{q\bar{q}}})] + K \\ &= \sum_{(q;\bar{q})=(1;1)}^{Q\bar{Q}} \frac{T_{q\bar{q}}}{T} m_{Y,q\bar{q}} + K \end{aligned}$$

Les moyennes sous les deux hypothèses sont:

$$\begin{aligned} m_{X,q\bar{q}} &= (\mu_q - \mu_{\bar{q}})^+ \Gamma_{\bar{q}}^{-1} (\mu_q - \mu_{\bar{q}}) \\ &\quad + tr(\Gamma_{\bar{q}}^{-1}\Gamma_q - I) \\ m_{\bar{X},q\bar{q}} &= -(\mu_q - \mu_{\bar{q}})^+ \Gamma_{\bar{q}}^{-1} (\mu_q - \mu_{\bar{q}}) \\ &\quad + tr(I - \Gamma_q^{-1}\Gamma_{\bar{q}}) \end{aligned}$$

On procède de même pour les variances:

$$\begin{aligned} \sigma_Y^2 &= E_Y[(\hat{ll}(O) - E_Y[\hat{ll}(O)])^2] \\ &= \sum_{(q;\bar{q})=(1;1)}^{Q\bar{Q}} \frac{T_{q\bar{q}}}{T} \sigma_{Y,q\bar{q}}^2 \end{aligned}$$

et

$$\begin{aligned} \sigma_{X,q\bar{q}}^2 &= (\mu_q - \mu_{\bar{q}})^+ \Gamma_{\bar{q}}^{-1}\Gamma_q\Gamma_{\bar{q}}^{-1} (\mu_q - \mu_{\bar{q}}) \\ &\quad + tr((\Gamma_{\bar{q}}^{-1}\Gamma_q - I)^2) \\ \sigma_{\bar{X},q\bar{q}}^2 &= (\mu_q - \mu_{\bar{q}})^+ \Gamma_q^{-1}\Gamma_{\bar{q}}\Gamma_q^{-1} (\mu_q - \mu_{\bar{q}}) \\ &\quad + tr((I - \Gamma_q^{-1}\Gamma_{\bar{q}})^2) \end{aligned}$$

3. EXPÉRIMENTATIONS

3.1. Contexte

Les expériences ont été réalisées avec la base de données SESP composée de séquences de chiffres néerlandais prononcés au téléphone dans des conditions réelles. Cette base est composée de 24 hommes et 24 femmes appellant de différents combinés et environnements. Lors de chaque communication, il a été demandé aux locuteurs de produire des séries de chiffres dépendants et indépendants de leur identité.

Pour initialiser les paramètres des modèles de locuteur, on utilise seulement 2 sessions d'apprentissage. Celles-ci correspondent à des appels avec un faible niveau de bruit, passés sur deux combinés différents. Le modèle de non-locuteur est commun à tous les clients et initialisée avec une base de locuteurs différentes de celle des locuteurs clients (modèle du monde). L'ensemble de test est composé de 1658 tentatives d'accès de clients et 486 d'imposteurs.

La paramétrisation acoustique choisie est composée de 16 coefficients de cepstre LPC, du logarithme de l'energie, ainsi que de la dérivée première et seconde de ces coefficients. Une soustraction cepstrale est appliquée aux paramètres.

Deux types de modèles de locuteur sont testés. Un HMM ergodique de quatre états et un HMM gauche droite de 5 états par phonèmes. Le nombre de fonctions gaussiennes pour chaque état est égal à 1. Lors des tests, les topologies des modèles des locuteurs et du non-locuteur sont identiques. Un modèle de silence, commun aux deux modèles, permet d'éliminer la contribution des trames correspondant aux silences ou aux bruits.

3.2. *Evaluation*

Pour nos expériences, nous avons choisi de supposer un équilibre des coûts et des probabilités des tentatives d'accès d'imposteurs et des clients. Le taux

$$C[\tilde{X} \mid \bar{X}].p_{\bar{X}} + C[\bar{\tilde{X}} \mid X].p_X$$

est choisi normalisé à 1 et les probabilités d'occurence des hypothèses sur les locuteurs sont définies par:

$$p_{\bar{X}} = \frac{C[\tilde{X}|X]}{C[\tilde{X}|X]+C[\bar{X}|\bar{X}]} \quad p_X = \frac{C[\bar{X}|\bar{X}]}{C[\tilde{X}|X]+C[\bar{X}|\bar{X}]} \qquad (12)$$

Ceci nous donne une fonction totale d'erreur estimée égale à:

$$\hat{R}_{dcf} = \frac{1}{2}\left[\hat{P}_{\bar{X}}(\tilde{X}) + \hat{P}_X(\bar{\tilde{X}})\right] \qquad (13)$$

et un seuil de décision théorique $\theta = 1$. Cette valeur correspond aussi au demi-TER (Half Total Error Rate).

Les quantités $\hat{P}_{\bar{X}}(\tilde{X})$ et $\hat{P}_X(\bar{\tilde{X}})$ correspondent respectivement aux fréquences de fausses acceptations (FA) et de faux rejets (FR). Elles sont mesurées tous locuteurs confondus.

3.3. *Résultats*

Trois expériences sont présentées avec les deux modèles HMM. La première (a) consiste à prendre la décision à la sortie de la modélisation des observations (2) avec la valeur de seuil théorique ($\theta = 1.0$). Elle sert de référence. Pour la seconde expérience (b), on prend la décision en suivant l'hypothèse de distribution gaussienne pour le logarithme de la vraisemblance $\hat{ll}$ (10). Enfin pour la dernière (c), on suppose une égalité des variances σ_Y, ce qui revient à ne prendre que les moyennes estimées. Cette dernière expérience a été motivée par les résultats de la précédente. Tous ces résultats se retrouvent dans la table 1.

Table 1: Scores de vérification

méthode	HMM					
	ergodique			gauche-droite		
	FR	FA	$\hat{R}_{dcf}$	FR	FA	$\hat{R}_{dcf}$
(a)	13.3	2.6	7.9	65.0	0.0	32.5
(b)	5.8	9.4	7.6	15.4	0.4	7.9
(c)	7.1	4.8	5.9	7.3	1.0	4.1

On constate une amélioration du résultat de la décision lors de l'utilisation d'un modèle de distribution de la vraisemblance. Cette amélioration est beaucoup plus nette pour le modèle gauche-droite que pour le modèle ergodique. Cette différence peut s'expliquer par un partitionnement beaucoup plus fin de l'espace des observations et l'ajout d'une contrainte sur l'ordre chronologique des trames.

Ne pas tenir compte des variances améliore les résultats. Une explication possible est la non adéquation de nos hypothèses pour son calcul. Lorsque on estime $\sigma_Y^2 = E_Y[(\hat{ll}(O) - E[\hat{ll}(O)])^2]$, des moments d'ordres 3 des observations interviennent et ne sont pas forcément nul comme le suppose les hypothèses des Chaînes de Markov Cachées. Une autre explication peut être le nombre peu élevé d'échantillons attribué à certains états et qui remet en cause les hypothèses de convergence vers une gaussienne de la distribution de la vraisemblance.

4. CONCLUSION

Cet article est un pas supplémentaire dans la description des étapes de modélisation en vérification du locuteur. Il permet ainsi de rappeller l'importance de la prise en compte des approximations successives, des hypothèses et des aspects analytiques effectuées lors de l'estimation du seuil de décision. On montre aussi que la recherche d'une description précise du processus de vérification peut apporter de nettes améliorations.

Ce travail ne tient pas compte de l'inadéquation qui existe entre les données d'apprentissage et de test. Mais cet aspect n'a volontairement pas été traité pour éviter la confusion des effets sur le score final. Le prolongement de cette étude doit porter maintenant sur la correction de ce biais.

BIBLIOGRAPHIE

[Sch90] L.L. Scharf : *Statistical Signal Processing: detection, estimation, and time series analysis.* Addison Wesley. 1990

[Fuk72] K. Fukunaga : *Introduction to statistical pattern recognition.* Academic Press. 1972.

[Bim96] F. Bimbot, H.P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, J.B. Pierrot : *An overview of the CAVE project research activities in speaker verification.* Eurospeech 97. 1996.

[Pie98] J.B. Pierrot, J. Lindberg, J. Koolwaaij, H.P. Hutter, D. Genoud, M. Blomberg, F. Bimbot : *A comparison of a priori threshold setting procedures for speaker verification in the CAVE project.* ICASSP 98. 1998.

[Bim97] F. Bimbot, D. Genoud : *Likelihood ratio adjustment for the compensation of model mismatch in speaker verification.* Eurospeech 97, pp 1387-1390. 1997.

[You97] S. Young, J. Jansen, J. Odell, D. Ollason, P. Woodland : *The HTK Book, HTK 2.1* Manual. 1997

[Sap90] G. Saporta : *Probabilités. Analyse des données statistiques.* Livre, Technip-Paris. 1990

Proposition d'une stratégie de fusion de données à trois niveaux pour la vérification d'identité

*Patrick Verlinde** *Dominique Genoud*† *Guillaume Gravier* *Gérard Chollet*

Ecole Nationale Supérieure des Télécommunications / Département Signal
46, rue Barrault
75634 Paris Cedex 13, FRANCE

ABSTRACT

The aim of this paper is to propose a strategy that uses data fusion at three different levels to gradually improve the performance of an identity verification system. In a first step temporal data fusion can be used to combine multiple instances of a single (mono-modal) expert to reduce its measurement variance. If the system performance is too low, one can try to improve it by fusing in a second step the results of multiple experts working on the *same* biometric modality. This supposes that their respective classification errors are (at least partially) uncorrelated. Finally, if the performance is still not good enough, one will be forced to pass on to the third step in which performance can be increased when using multiple experts working on different (biometric) modalities. These experts have to be chosen in such a way that the effect of adding the extra modalities is to increase the separation between the distributions of the different populations that have to be classified by the system. This kind of strategy allows a user to gradually tune the performance of an identity verification system to its needs while controlling the increase of investment costs[1].

1. INTRODUCTION

L'objectif d'un système de vérification d'identité est de décider si une personne qui proclame une identité est bien la personne qu'elle prétend être. Chaque méthode implémentée, appelée ici expert, effectue un test d'hypothèse binaire. Ces tests génèrent deux types d'erreurs: un faux rejet (FR: rejeter un accès client) ou une fausse acceptation (FA: accepter un imposteur). Les performances d'un système de vérification d'identité sont en général caractérisées par les taux d'erreurs globaux mesurés pendant les tests: le taux de faux rejets [False Rejection Rate: FRR = (nombre de FR) / (nombre d'accès client)] et le taux de fausses acceptations [False Acceptance Rate: FAR = (nombre de FA) / (nombre d'accès imposteur)]. Une mesure unique peut être obtenue en combinant ces deux erreurs pour obtenir le taux d'erreurs global [Total Error Rate: TER = (nombre de FA + nombre de FR) / (nombre total d'accès)] ou par son complément, le taux de réussite global [Total Success Rate: TSR = 1 - TER] [BC97]. Une vérification d'identité parfaite, FAR = 0 et FRR = 0, peut difficilement être obtenue en pratique. Cependant, une étude détaillée des tests d'hypothèse binaire [Tre68] montre que si l'on fait tendre FRR vers zéro en modifiant le seuil de décision, on augmente alors le taux de fausses acceptations, et inversement.

Les erreurs sont causées par l'imperfection des différents experts. On a donc des erreurs de mesure. Une solution, détaillée dans la section 2, pour remédier à ce problème consiste à combiner les résultats obtenus par un seul expert sur des instances successives sur plusieurs tests. On parlera alors de fusion temporelle. De cette manière on peut effectivement diminuer la variance des résultats, ce qui permet d'augmenter les performances du système de base sans avoir à modifier l'approche choisie. Si ces performances ne satisfont pas les besoins de l'utilisateur, il peut passer à l'amélioration suivante.

Le fait de disposer d'experts parfaits et de connaître les distributions statistiques des résultats obtenus par chaque méthode de vérification pour les deux classes (clients, imposteurs), permet d'utiliser la règle de décision bayésienne pour obtenir un fonctionnement du système au taux d'erreur minimal [DH73]. L'erreur de classification ne sera égale à zéro que dans le cas où les deux distributions ne se recouvrent pas du tout. Si cette condition n'est pas satisfaite, ce qui est le cas en général, on ne peut faire baisser les deux types d'erreurs (FAR, FRR) *simultanément* qu'en augmentant la taille de l'échantillon statistique, c'est-à-dire en utilisant plus de données d'entraînement, permettant ainsi de diminuer les variances des estimateurs des paramètres réels de la distribution. Très souvent on ne dispose malheureusement que d'un nombre limité de données d'entraînement, ce qui exclut cette possibilité théorique. Une autre possibilité pour faire diminuer ces deux taux d'erreurs simultanément consiste alors à fusionner les résultats de différents experts basés sur la même modalité. Ce type de fusion permet d'augmenter les performances si les erreurs commises par les experts ne sont pas totalement corrélées. L'amélioration des performances sera même d'autant plus grande que les résultats (et les erreurs) obtenus par les différents experts seront plus décorrelés, l'apport d'information augmentant avec la décorrelation. Dans la section 3, on montrera le cas d'un système monomodal multi-expert basé sur la modalité biométrique vo-

*Ecole Royale Militaire / Département ELTE, Bruxelles, Belgique, e-mail: patrick.verlinde@tele.rma.ac.be

†Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Suisse, e-mail: genoud@idiap.ch

[1]Some of the results given in this paper were obtained in the framework of the CAVE-PICASSO and M2VTS European projects.

cale. Dans le cas spécifique d'une modalité vocale, on a ajouté la possibilité de différer la décision après fusion en cas de doute. Cette technique permet de diminuer les erreurs au prix d'une augmentation du nombre de non-décisions et d'un traitement supplémentaire. Ce mécanisme peut également être utilisé pour les autres systèmes de vérification d'identité, mais cette augmentation des performances induit des coûts supplémentaires. Si les performances du système après fusion mono-modale ne satisfont toujours pas l'utilisateur, il est possible de procéder à l'étape suivante.

Dans un troisième temps on peut espérer augmenter encore les performances en trouvant des caractéristiques qui améliorent la séparabilité des hypothèses testées en augmentant la dimension de l'espace. Il est possible d'utiliser à cette fin des modalités biométriques supplémentaires. La discrimination entre les deux distributions sera d'autant plus facile que la correlation entre les différentes modalités sera plus faible. On étudiera ce cas dans la section 4 pour des modalités biométriques vocales et visuelles.

2. FUSION MONO-MODALE MONO-MÉTHODE

2.1. *Méthodes utilisées*

Le premier pas dans la stratégie proposée, est d'utiliser une intégration temporelle afin d'augmenter les performances d'un système de vérification d'identité en diminuant la variance des mesures. Un exemple de cette technique est présenté dans [KMJS97]. Les auteurs de cet article partent, d'une part, du principe que les décisions d'accepter (ω_1) ou de rejeter (ω_2) la personne, sont basées sur la probabilité de classes *a posteriori* $P(\omega_j|x_i), j = 1,2$ et d'autre part, du fait que l'on dispose de plusieurs instances x_i de données biométriques. Dans ce cas, on peut se baser sur la fusion de l'information additionnelle. On suppose que les mesures x_i ont été acquises dans les mêmes conditions et qu'elles peuvent être considérées comme des mesures multiples qui ne diffèrent entre elles que par leur composante de bruit. Aussi les probabilités de classes *a posteriori* peuvent être considérées comme des estimations bruitées de la valeur nominale de cette probabilité, *i.e.*:

$$P(\omega_j|x_i) = P(\omega_j|x) + \epsilon(\omega_j|x_i).$$

Ainsi on peut obtenir de meilleures estimations $\hat{P}(\omega_j|x)$ de cette probabilité *a posteriori* en combinant les estimations bruitées de manière linéaire ou en utilisant des statistiques de rang. Les différentes stratégies de fusion sont comparées entre elles en analysant leur probabilité d'erreur de manière théorique. Les méthodes de fusion étudiées sont les suivantes:

Moyenne arithmétique

$$\hat{P}(\omega_j|x) = \frac{1}{R}\sum_{i=1}^{R} P(\omega_j|x_i).$$

Règle maximale

$$\hat{P}(\omega_j|x) = \max_{i=1}^{R} P(\omega_j|x_i).$$

Règle minimale

$$\hat{P}(\omega_j|x) = \min_{i=1}^{R} P(\omega_j|x_i).$$

Règle médiane

$$\hat{P}(\omega_j|x) = \text{med}_{i=1}^{R} P(\omega_j|x_i).$$

2.2. *Résultats*

On démontre que la réduction de la variance est la plus spectaculaire pour la moyenne arithmétique près de la frontière de décision. Les réductions de la variance pour les statistiques de rang dépendent, entre autres, du nombre d'instances R et de l'ordre du rang et, bien qu'elles soient moins efficaces que dans le cas de la moyenne arithmétique, elles sont plus robustes par rapport aux valeurs extrêmes des résultats de mesure. La table 1 montre une réduction des taux d'erreurs jusqu'à 40%. Le gain en performance est monotone au début d'une séquence, mais tend à saturer après intégration des premières observations. Ces résultats ont été obtenus en utilisant un expert de vérification visuelle, basé sur une correlation robuste optimisée. Pour ces experiences, l'expert utilise uniquement la modalité image de la base de données multimodale M2VTS [PV97] qui contient 37 personnes.

Table 1: EER en fonction du nombre d'instances

Expert	1	2	3	4	5	6
Moyenne	6.9	4.6	4.5	3.7	4.0	4.0
Médian	6.9	4.8	4.2	4.3	4.2	4.0

3. FUSION MONO-MODALE MULTI-MÉTHODE

3.1. *Méthodes utilisées*

La combinaison de plusieurs experts sur une même modalité biométrique est très utilisée. On peut citer par exemple [Ben92, XKS92].

A titre d'exemple plus élaboré, les experiences décrites dans [GBGC96], basées sur plusieurs méthodes de vérification du locuteur en mode dépendant du texte, sont decrites. les trois experts utilisés acceptent en entrée un ensemble de vecteurs de 12 coefficients cepstraux LPC extraits d'un signal de parole. Ces experts sont:

1. *Dynamic Time Warping (DTW)* L'algorithme de DTW consiste à effectuer une comparaison dynamique entre une matrice de référence et une matrice de test. Le résultat est une mesure de distance entre le test et la référence.
2. *Statistique du second ordre (SSO)* Une matrice de covariance $X = \frac{1}{M}\sum_{t=1}^{t=M} X_t X_t^T$ de la séquence de parole de référence est créée. Une matrice de covariance Y de la séquence de test est également créée. Une mesure de sphéricité symétrique $\mu_{AH}(X,Y)$ est ensuite effectuée.

3. *Modèles de Markov cachés (HMM)* Deux sortes de modèles HMM sont créés pour chaque mot à tester: (1) **Un modèle du monde**, créé à partir d'une base de données (Polyphone [CCCL95]). Les modèles de monde sont identiques pour tous les clients. (2) **Un modèle du client**, qui utilise comme paramètres initiaux le modèle du monde, et dont on réestime les paramètres pour chaque locuteur avec les données de celui-ci. La mesure utilisée consiste à calculer le log du rapport de vraisemblance locuteur/monde pour un segment de parole donné.

Détermination des seuils La décision d'acceptation ou de rejet d'un locuteur s'effectue par comparaison du score de chaque methode à un *seuil a priori*. Le seuil est déterminé par la méthode de Furui. Ce seuil est robuste lorsque très peu de données d'entraînement sont à disposition [Fur81].

Combinaison des décisions Afin d'améliorer la décision globale de notre système, nous avons combiné les décisions de chaque méthode (DTW, SSO, HMM). La combinaison utilisée ici est un système majoritaire pondéré avec seuil de doute. La décision par rapport au seuil *a priori* est pondérée (entre 0 et 1) en fonction de la distance seuil/score, ce qui revient à établir une confiance dans cette décision. Le système compare la moyenne des confiances des méthodes qui ont obtenu la majorité à un seuil (seuil de doute). Si cette moyenne est supérieure au seuil, on prend la décision de la majorité, sinon il y a doute.

3.2. *Résultats*

La base de données Polycode [GC95] utilisée pour ces expériences est constituée de 25 locuteurs. La table 2 montre que la fusion des décisions des trois experts est meilleure que le meilleur des experts. Pris dans un cadre plus formel, les trois experts utilisés ne sont pas fondamentalement différents, la DTW et les SSO étant des cas "dégénérés" de la modélisation HMM. Cependant la modélisation des locuteurs effectuée par chaque expert n'étant pas parfaite, ce sont parfois les algorithmes nécessitant moins de paramètres à estimer qui s'avèrent les plus robustes. Parmi les facteurs qui entachent l'estimation correcte d'un modèle, le manque de données d'entraînement et la variabilité intra-locuteur sont certainement les plus importantes.

Table 2: Performance des experts et de la fusion

Expert	FRR% (200 tests)	FAR% (1800 tests)
DTW	23.5	7.67
SOSM	14.0	5.28
HMM L/R	5.53	2.72
Fusion	2.0	2.72

Malgré l'amélioration des performances due à la fusion, les erreurs résiduelles peuvent, dans certaines applications, s'avérer inacceptables, et il est possible, dans ce cas, d'éviter de prendre de mauvaises décisions en rejetant les cas douteux. La table 3 nous indique le comportement des performances du système lors du changement de seuil de doute. Ce seuil permet de régler ainsi les taux de faux rejets et fausses acceptations selon le niveau de sécurité requis par l'application. Les cas douteux peuvent être résolus en demandant des informations supplémentaires au locuteur.

Table 3: Performance en changeant le seuil de doute

Seuil de doute	FRR% (200 tests)	FAR% (1800 tests)	Doute% (2000 tests)
0.2	2.0	2.72	0.0
0.5	2.0	2.33	0.83
0.7	2.0	1.11	10.11
0.8	1.0	0.89	20.2

4. FUSION MULTI-MODALE MULTI-MÉTHODE

4.1. *Méthodes utilisées*

Le troisième pas dans la stratégie de développement proposée est la fusion de multiples experts utilisant plusieurs modalités (biométriques). Cette approche a, entre autre, été utilisée dans les références suivantes [BF95, DBB+97, JLGW97].

Pour montrer les possibilités de la fusion multi-modale, nous avons testé différents modules de fusion et nous avons comparé les résultats à ceux obtenus avec les experts individuels, sur la base de données multi-modale M2VTS [PV97] avec cinq experts:

UCL Pro expert visuel basé sur les profils (Université Catholique de Louvain);

UCL Grey expert visuel basé sur les profils et les niveaux de gris (Université Catholique de Louvain);

RMA Pro expert visuel basé sur les profils (Ecole Royale Militaire);

UOS Cor expert visuel frontal basé sur une méthode de corrélation robuste (University of Surrey);

RMA Ar expert vocal basé sur une méthode autorégressive (Ecole Royale Militaire).

On a testé les modules de fusion suivants: méthodes par vote (ET, OU, Majoritaire) [Das94], classificateurs linéaire et quadratique [DH73] et le perceptron multicouches (MLP) [Bis95].

4.2. *Résultats*

Les résultats expérimentaux se trouvent dans la table 4. Ces résultats, exprimés en taux d'égales erreurs, ont été obtenus en utilisant les seuils *a posteriori* et représentent donc les meilleures performances possibles.

Après ce premier test, on a comparé la meilleure méthode avec un MLP (caractérisé par 25 neurones logsig dans la

Table 4: Taux d'égales d'erreurs avant et après fusion

Expert	EER%
UCL Pro	08.0
UCL Grey	07.0
RMA Pro	17.0
UOS Cor	12.0
RMA Ar	02.0
Méthode par vote ET	12.5
Méthode par vote OU	06.5
Méthode par vote MAJ	05.0
Classificateur linéaire	02.5
Classificateur quadratique	01.0

couche cachée et un neurone logsig dans la couche de sortie), sur des données de test en utilisant les seuils calculés *a posteriori* (selon le critère EER). Les résultats se trouvent dans la table 5. On constate que la fusion de données multi-modale permet effectivement d'augmenter les performances au delà des performances du meilleur expert.

Table 5: Comparaison des erreurs de vérification

Méthode	FRR%	FAR%
Classificateur quadratique	0.0	4.1
MLP	0.0	0.4

5. CONCLUSIONS

La fusion de données peut être utilisée au sein d'un système de vérification d'identité à différents niveaux. Le premier niveau est la fusion temporelle des résultats obtenus par un seul expert, et donc sur une seule modalité, afin de réduire la variance de mesure. Un deuxième niveau est atteint en fusionnant les résultats obtenus par différents experts sur la même modalité, ceci afin de minimiser les erreurs de classification en jouant sur la décorrelation des erreurs des différents experts. Le troisième niveau d'application est de diminuer encore les erreurs de classification en essayant d'accroitre la distance entre les distributions des deux populations en augmentant la dimension de l'espace par l'utilisation de plusieurs modalités (biométriques) les plus décorrelées possible. Nos travaux actuels sont axés vers une combinaison de ces trois utilisations de la fusion de données dans une seule stratégie séquentielle, qui a le grand avantage d'augmenter les performances d'un système de vérification d'identité de manière graduelle.

BIBLIOGRAPHIE

[BC97] F. Bimbot and G. Chollet. Assessment of speaker verification systems. In *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, 1997.

[Ben92] Y. Bennani. Text-independent talker identification system combining connectionist and conventional models. In S. Y. Kung et al., editor, *Neural Networks for Signal Processing, Vol.2*. IEEE Service Center Press, 1992.

[BF95] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford UK, 1995.

[CCCL95] G. Chollet, J.L. Cochard, A. Constantinescu, and Ph. Langlais. Swiss french polyphone and polyvar: telephone speech databases to study intra and inter speaker variability. Technical report, IDIAP, 1995.

[Das94] B. V. Dasarathy. *Decision Fusion*. IEEE Computer Society Press, 1994.

[DBB+97] B. Duc, E. Bigün, J. Bigün, G. Maître, and S. Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):835–843, 1997.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[Fur81] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on ASSP*, ASSP-29(2):254–272, 1981.

[GBGC96] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet. Combining methods to improve speaker verification decision. In *Proceedings of The Fourth International Conference on Spoken Language Processing*, 1996.

[GC95] Dominique Genoud and Gérard Chollet. Polycode a verification database. Technical report, IDIAP, CH-1920 Martigny, 1995.

[JLGW97] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):845–852, 1997.

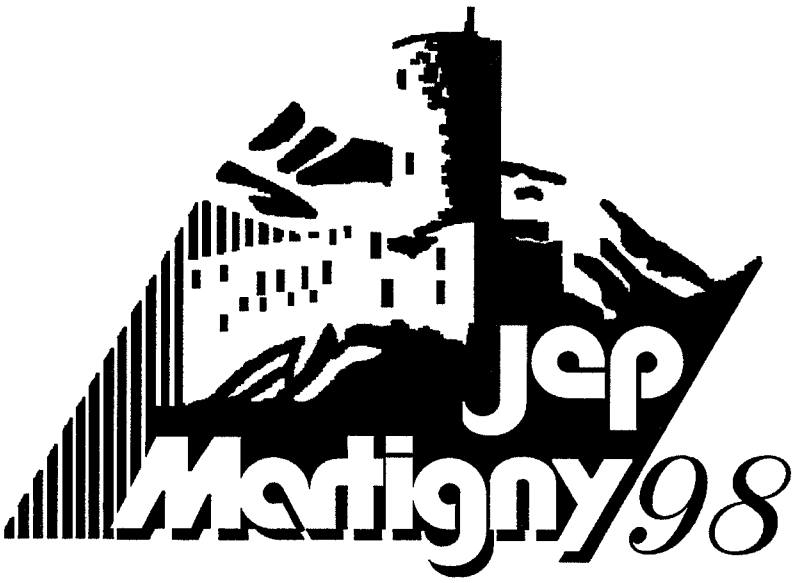
[KMJS97] J. Kittler, G. Matas, K. Jonsson, and M. U. Ramos Sánchez. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18(9):845–852, 1997.

[PV97] S. Pigeon and L. Vandendorpe. The M2VTS multi-modal face database. In *Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, Lecture Notes in Computer Science. Springer Verlag, 1997.

[Tre68] H. L. Van Trees. *Detection, Estimation and Modulation Theory*, volume 1. John Wiley & Sons, New York, 1968.

[XKS92] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, May/June 1992.

Psycholinguistique



Les opérations pragmatiques d'une communication agrammatique

Françoise Argod-Dutard

Université Michel de Montaigne-Bordeaux III
33405 Talence Cédex, Fax 05 56 84 51 30, tél. 05 57 87 27 82.

ABSTRACT

Discourse theory and pragmatics are used for a quantitative and a qualitative analysis of agrammatic speech in a case of Broca's aphasia. Results evince both global and selective deficiencies in the syntax : intanciation and referenciation are severely impaired; the patients's speech is of low locutive power, and the implicit content is increased. Agrammaticism is a disorder in enunciation which results from a communicational economy in speech act.

1. INTRODUCTION

Les nombreuses études neurolinguistiques qui ont été consacrées jusqu'à présent à l'agrammatisme portent essentiellement sur l'analyse des signifiants, sur celle de leurs possibilités combinatoires et génératives dans le cadre de la phrase, sur celle de la fonction informative du langage agrammatique.
Les résultats qui, dans le premier cas, correspondent à l'application de méthodes fonctionnalistes comme celle de A. Martinet, ont été obtenus à l'aide de protocoles d'examens métalinguistiques ou d'entretiens dirigés [TIS73]. L'agrammatisme se définirait par une rareté des morphèmes grammaticaux et donc une prédominance des lexèmes substantivaux non marqués, la fréquence des verbes à infinitif; la perturbation de la prosodie phrastique, la réduction des énoncés à un ou deux mots ou la non conservation de la structure ordonnée des syntagmes sont des éléments fréquents mais non unanimement reconnus. La deuxième approche qui utilise les outils analytiques de la grammaire générative a montré à l'aide de protocoles métalinguistiques, de tests inducteurs, que l'agrammatique se fonde, dans ses jugements à propos des liens des unités du syntagme et de la phrase, sur un schéma hiérarchique excluant les faits non essentiels et les propriétés de surface, qu'il préfère les formes lexicales et syntaxiques non marquées et fréquentes, et serait influencé par la phonologie des séquences élémentaires de la phrase pour l'omission de certaines marques morphophonologiques [McE95,GOO97]. L'application de la théorie de la communication a démontré, grâce à l'entretien dirigé, à la comparaison d'agrammatismes de Broca, d'agrammatismes d'enfants et d'étrangers notamment, que le processus linguistique était le même et reposait sur des principes de simplification qui varient selon la complexité et l'ambiguïté de la communication (déficit de la performance, difficultés de stratégie, theory of trace deletion) [NES79,MAH95,GRO95]. Mais il reste que ces approches sont limitées au cadre de la phrase et aux structures de surface, qu'elles constatent des variations quantitatives dans la fréquence des signifiants du code sans pouvoir les expliquer autrement que par une hypothèse générale de déficits linguistiques. Or la variabilité d'occurrences identiques pose problème. Quelles sont les règles qui décident, par exemple, que certains articles sont omis ici et non là, que certains pronoms personnels existent alors qu'on cherche les autres, qu'un agrammatique utilise tout de même la flexion ? En somme, quelles sont les opérations mentales complexes sous-jacentes dont les signifiants ne seraient que les marques ? Le recours au sens et au rôle des unités dans la communication semble donc s'imposer pour essayer d'avancer dans l'approche qualitative de l'agrammatisme vers un modèle interprétatif de certains de ses aspects. C'est ce que nous avons essayé de réaliser en appliquant les concepts de la linguistique pragmatique à l'étude d'un langage agrammatique.

2. MATERIEL ET METHODE

2.1 La linguistique discursive et pragmatique

La théorie de l'énonciation, appropriation de la langue par un acte individuel d'utilisation, a été formulée en 1966 par E. Benveniste et précisée ensuite par tout un courant de linguistes bien connus qui, en se penchant sur l'étude des présupposés énonciatifs, ont montré que toute prise de parole entrait dans une pragmatique reposant sur un certain nombre de fonctions discursives qui mettent en jeu parole, langue et situation . En effet, l'acte de langage est inscrit dans une situation générale, présupposée, contextuelle et énonciative : accompli par un locuteur, il comporte un énoncé reflétant diverses structurations, s'adresse à un destinataire qui devra synthétiser toutes les variables langagières et comporte une efficience par rapport à l'ensemble de la situation. Si l'on considère l'énoncé, on constate que ses structurations sont produites essentiellement par une dizaine d'opérations qui, en utilisant des marques —prépositions, ordre des mots, accords, prosodie et morphèmes —donnent naissance à un ensemble de plus en plus complexe du point de vue syntaxique et, par conséquent, du point de vue sémantique : le signifié dépend du sens propre ou contextuel des lexèmes mais aussi de leurs connexions, de leurs catégories et de leurs présupposés. **Les fonctions paradigmatiques** du langage regroupent des opérations de sélection qui jouent sur un axe syntagmatique (il s'agit alors d'itération, par addition, répétition, alternative, interrogation, élimination), soit sur deux axes (il s'agit alors d'insertion, d'apposition ou de discours rapporté). **Les fonctions de prédication** permettent d'attribuer une propriété particulière à un élément donné. Les opérations de caractérisation et de pluralité, qui interviennent après référenciation absolue, donnent un genre et un nombre aux unités lexicales choisies, la quantification, la

spécification qualitative et l'assertion déterminent les choix et les mettent en relation avec la logique. Elles s'appuient sur des opérations de connexion qui, en assurant des liens de dépendance, complexifient le sens du message.**Les fonctions de repérage** situent l'énoncé par rapport à la situation d'élocution présente ou au contexte large. Par l'instanciation, le locuteur s'approprie l'appareil formel de la langue et ce, par le biais de la catégorisation et de la déixisation. Avec la catégorisation nominale, il attribue la définitude ou la non définitude au substantif dans une situation d'énonciation présente. Par la catégorisation verbale, il donne au verbe ses marques de temps, d'aspect et de mode relativement à un repère temporel. La déixisation lui permet d'organiser son discours à partir de sa propre situation d'énonciation.Quant à la référenciation, c'est une opération complexe qui sert à appeler sur le même axe syntagmatique un même élément plusieurs fois à l'aide de substituts et à le situer par rapport au contexte. **Les fonctions de modalisation** donnent la possibilité au locuteur d'exprimer des attitudes variées à l'égard de son énoncé:modalisation de communication, affective, logique ou pragmatique. **Les fonctions de communication et d'argumentation** consistent à faire entrer les opérations précédentes dans une stratégie d'échange : ce sont les thématisations et les focalisations qui sont plus particulièrement concernées. L'étude de la qualité et de la quantité de ces opérations dans le discours agrammatique nous a donc paru propre à montrer comment l'acte linguistique était perturbé dans ses composantes et quels déficits devait essayer de combler la rééducation.

2.2 Le cas clinique

La patiente est une française née au Maroc, âgée de 63 ans, qui a présenté à l'âge de 40 ans une thrombose sylvienne gauche complète suivie d'une hémiplégie droite massive avec aphasie. Après une semaine de coma, elle a présenté une suspension totale du langage avec perte des automatismes et apparition de stéréotypies. La compréhension était bonne mais l'expression inexistante. Après un an de rééducation, elle est arrivée à un niveau de langage, qui est resté inchangé depuis, caractérisé par une perte de l'écrit, des troubles d'expression syntaxique et une compréhension verbale autorisant néanmoins la transcription non équivoque de ses propos.

2.3. Les procédés d'enquête et d'analyse

Pour recueillir des échantillons de langage oral spontané, et représentatif, nous avons sélectionné, à partir de conversations dirigées, quatre types de discours : le récit d'une journée à la maison et d'une journée professionnelle, une discussion sur le travail de la femme et une narration sur images séquentielles. Les enregistrements ont été faits à domicile et représentent une heure et quart. Après avoir mis en rapport les 893 mots recueillis avec les 1110 mots du langage normal de sa soeur, nous avons recouru ensuite à sa structure restituée pour isoler les opérations et les marques présentes et absentes ou perturbées. Cette approche a ainsi permis de pratiquer une étude comparative à deux niveaux : langage restitué et langage témoin. Le relevé des marques et des opérations nous a conduite à procéder alors à des pourcentages relatifs à l'ensemble ou restreints au cadre d'une seule opération pour les trois niveaux de langage et les quatre types de discours.

3. LES RESULTATS

Si l'on considère la fig. 1 qui représente toutes les opérations étudiées dans le détail, on constate qu'il y a réduction générale du nombre des opérations de 40,86% par rapport au langage restitué et 45,70% par rapport au langage normal. Seules **la modalisation** et **la quantification** sont sauvegardées avec ces particularités que la première est surtout affective et pragmatique, la seconde plus objective que subjective. Trois opérations, la prédication assertive, les opérations paradigmatiques et de mise en relief, perdent de leurs qualités. Il y a eu, dans **l'assertion**, évasion des verbes et des opérateurs et augmentation parallèle des autres éléments prédicatifs tandis que les marques conjointes d'accord ont diminué également, dans la prédication complexe, perte des conjonctions, dans la fonction impérative, baisse du nombre des verbes. La prédication a donc tendance à devenir, de verbale, nominale ou adjectivale, le sens étant porté plus par l'élément adjoint que par l'opérateur comme en témoignent les exemples suivants : "comprimés impossible" (pour "sans comprimés, c'est impossible"), "Rabat, marché" (pour "à Rabat, il y a le marché"). **Les opérations paradigmatiques** sont marquées plus par l'intonation que par des morphèmes, ce qui leur confère un caractère moins explicite et clair que pour celles du parler témoin, comme l'illustrent les phrases suivantes : "travailler le soir ménage manger" (pour "elle travaille le soir pour faire le manger et le ménage"). Dans **les opérations de mise en relief**, seule, la marque constituée par le pronom de reprise du thème dans le rhème s'affaiblit. Notre patiente dit, par exemple, "moi, travailler ménage", "mon mari retraite maintenant" (pour "moi je travaille au ménage", "mon mari, il est à la retraite maintenant"). Si ces atteintes ne sont pas négligeables, constatons qu'elles sont partielles et n'atteignent que les opérations en nombre restreint. C'est dire que la réduction est sélective et frappe surtout les plus utilisées : caractérisation et pluralité, connexion instanciation et référenciation. **L'instanciation**, en particulier, est très atteinte. La catégorisation nominale diminue en extension et en qualité, les pertes atteignant les deux marques mais principalement l'indéfinitude. La catégorisation verbale s'affaiblit de façon générale: l'emploi des modes nominaux moins différenciés se généralise. La déixisation souffre dans le fonctionnement des marques des allocutaires, marques amalgamées et atones. Des phrases du type "soigner pansements plaies graves" (pour "je soignais avec des pansements des plaies graves") sont fréquentes chez notre patiente. L'étude de **la référenciation** fait apparaître une perte très nette d'anaphoriques et une diminution de plus de moitié du nombre des référenciations relationnelles. Les marques les plus atteintes sont les pronoms personnels, les démonstratifs et les accords conjoints, comme en témoignent les exemples "mort maintenant Docteur Dupont", "Monsieur agents expliquer comment arriver" (pour "Le monsieur aux agents explique comment c'est

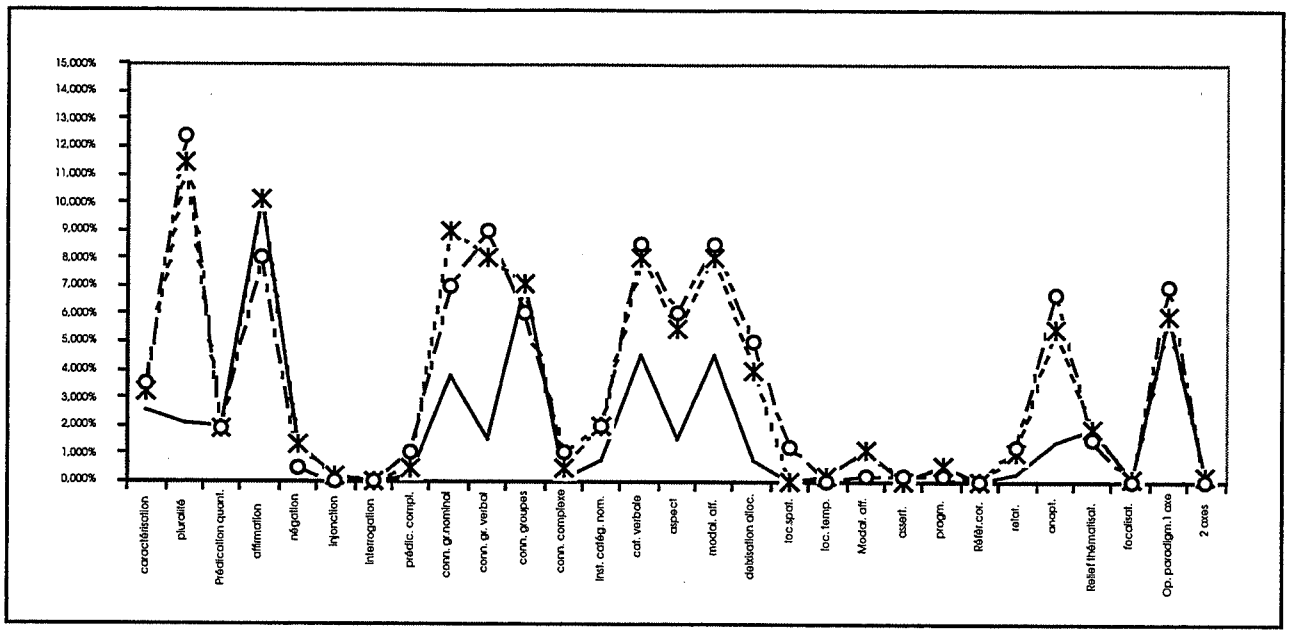


Figure 1 : Langage pathologique —— Langage restitué — O — O — Langage témoin ---x---x---

arrivé"). L'analyse de la connexion a montré que Madame R. utilisait prioritairement comme marques la prosodie et l'ordre des mots et qu'elle négligeait les liens à l'intérieur des groupes nominaux et verbaux alors qu'elle les maintenait entre les groupes eux-mêmes. Elle dit ainsi "trouvé dans la maison" (pour "ils l'ont trouvée dans la maison").Les opérations de **caractérisation** et de **pluralité** ont tendance à perdre de la redondance, les marques conjointes d'accords étant portées par le déterminant (les verbes et les adjectifs diminuant) et même à s'effacer faute de traits distinctifs signifiants essentiels qui marquent le léxème ou son substitut. Ainsi la phrase "déjeuner, ampoule, vitamines" est indifférenciée de ces points de vue. En ce qui concerne **les marques**, de façon générale, elles voient se réduire leur fréquence par rapport à celles des deux autres langages. Les pertes concernent surtout les marques d'accord (16,73%) et les marques grammaticales (24,34%). Ce sont les plus complexes présentant un amalgame de traits comme les pronoms et les morphèmes verbaux qui disparaissent en plus grand nombre et dont la perte, par les autres disparitions qu'elles impliquent, est le plus préjudiciable au système des marques et à la vie des opérations. Quant à l'incidence du type de discours sur les productions sollicitées, il n'a pas été possible de la mettre en évidence.

A partir des données recueillies, nous avons essayé de dégager quelques rapports précisant mieux **la notion de grammaticalité** dans le langage témoin et le langage restitué et celle d'**agrammatisme** dans le langage de Mme R. La proportion des marques par opération montre que les opérations restantes sont claires mais le rapport du nombre des marques diminue et celui du nombre d'opérations augmente. Aussi peut-on dire, d'une part, que la proportion d'implicite croît puisque les marques diminuent quand les opérations sont en progression et, d'autre part, qu'il y a économie du nombre de mots et de marques pour une même fonction grammaticale, l'auditeur faisant les frais de cette économie. La syntaxe n'étant qu'un moyen d'enrichir la sémantique, nous avons voulu chercher quelles étaient les conséquences de l'agrammatisme de Mme R. sur **la valeur de son discours dans l'acte de langage**, en fonction de l'étude du posé, des présupposés et de l'implicite. Dans le premier cas, la sémantique de l'énoncé souffre de la perte des référenciations et des connexions. la force locutive de la patiente est médiocre et l'auditeur doit sans cesse suppléer aux lacunes de l'expression linguistique et se projeter dans son discours pour construire les références omises et les relations manquantes, entre les éléments explicites et les indices grammaticaux latents. De plus, la mise en relation de l'énoncé avec la situation d'énonciation se fait mal puisque la perte de la fonction d'instanciation par rapport au langage restitué atteint 14% pour une fonction qui occupe 28% de l'ensemble des opérations. En revanche, l'aptitude de notre patiente à faire partager à l'auditeur, par les opérations de communication et d'argumentation, un ensemble de présuppositions semble correcte, de même que sa capacité à user de la situation et à agir sur le langage, en dépit de l'absence de performatifs, si l'on en juge par le nombre des modalisations pragmatiques. Finalement, le non-explicite est constant chez Mme R., l'usage des infinitifs, d'aspect général et indéterminé, celui des articles définis en position de neutralisation, le nombre des verbes omis, les absences de marques déjà étudiées, tout cela crée une sémantique particulière de l'énoncé : l'auditeur perçoit l'implicite ou donne un contenu personnel à tel ou tel propos ou interprète telle ou telle implication permise par le canevas du message. Il en résulte une pluralité de sens et c'est ce jeu de reconstruction permanente qui rend sans doute fatigante la conversation avec un agrammatique que l'on ne connaît pas.

4. DISCUSSION

Ces résultats amènent quelques commentaires si on les compare à ceux des travaux antérieurs, toutes proportions gardées. En effet, nous remarquons que le premier trait couramment observé par la majorité des auteurs, à savoir l'absence de mots outils ne nous est pas apparu clairement

décisif dans la mesure où notre corpus comporte de telles marques. Tout au plus, peut-on dire qu'ils sont en nombre moins fréquent. Le 2e trait, c'est à dire la prédominance des lexèmes, se trouve confirmée par notre travail. Il en va de même pour la fréquence des infinitifs qui représentent le 3e trait définitoire de l'agrammatisme. Quant au 4e, qui fait référence à la suppression des marques et au style télégraphique, il corrobore la notion d'économie grammaticale que nous avons mise en évidence. Au total, les signifiants de la langue pathologique observée sont assez semblables à ceux que relèvent les publications. Ce que nous avons voulu montrer, c'est que les modifications observées pouvaient s'expliquer dans leurs variations par **des différences de valeurs significatives dans l'acte énonciatif**. Autrement dit, des absences de mêmes signifiants peuvent correspondre à des opérations syntaxiques différentes, par exemple définitude et référenciation pour l'article, et certaines de ces absences peuvent résulter d'une économie entre coût opératoire et besoin de signifier en situation pragmatique où auditeur, situation, efficience linguistique sont autant de paramètres jouant sur le locuteur et l'énoncé. Ainsi, on a pu constater que les opérations de thématisation et de focalisation qui correspondent à des fonctions de communication et d'argumentation, dirigées vers l'auditeur, sont assez bien sauvegardées sans doute en raison de leur structure binaire, de la clarté de la marque intonative et de l'importance que requiert l'auditeur dans la communication pragmatique. Les modalisations et la quantification, prégnantes sur le plan lexical et simples sur le plan syntaxique, demeurent inchangées. Les opérations paradigmatiques évoluent dans le sens d'une simplification des marques qui deviennnent surtout intonatives sans nuire à la communication, les prédications sont soumises également à un type d'économie dans la mesure où les marques se dirigent sur les éléments porteurs de sens —noms et adjectifs— et permettent ainsi à la communication de subsister, bien que les opérations soient atteintes sur le plan des marques. Si l'on considère des opérations comme la pluralité ou la caractérisation, on remarque que leur baisse est aisément compensée par la quantification, forme de pluralité, ou par le contexte large ou restreint partagé par le locuteur qui servira de système de suppléance, l'acte de langage étant un tout. L'évolution des opérations de connexion se fait de même dans le sens d'une simplification des marques à l'intérieur des groupes réduits, nominaux ou verbaux, aisément supplées par l'auditeur sans atteindre les marques inter-groupes porteuses de la signification de l'ensemble de la phrase. Quant à l'instanciation et à la référenciation, qui impliquent un ancrage situationnel ou référentiel complexe sur le plan linguistique, elles voient leur système de marques se désagréger tout en étant aidées par la proximité du contexte réel. **La transformation des marques répond au même principe.** L'intonation clairement perçue par l'auditeur suffit à assurer la compréhension des opérations , de même que les éléments lexicaux qui subsistent et l'ordre des mots. L'agrammatique se trouve donc, dans la communication, devant plusieurs difficultés à surmonter. S'il a la compétence linguistique du code, il n'a pas la maîtrise des opérations complexes : il recule devant les itérations paradigmatiques, la superposition des axes du discours, il n'aime pas enchâsser les prédications et les connexions, utiliser des marques amalgamées. A ces inhibitions de codage linguistique, s'ajoutent ses difficultés d'appropriation personnelle de la langue. S'impliquer subjectivement dans son discours, autrement que par l'intonation, lui est quasiment impossible. De plus, il est rebuté par les repérages successifs que nécessite tout message par rapport à son environnement textuel. Comment dès lors communiquer quand le discours doit se construire dans une perspective pragmatique, est ancré dans le temps, l'espace et un réseau de relations interindividuelles ? Il recourt à plusieurs solutions d'abord la simplification par l'ellipse, la réduction du nombre d'opérations par l'absence de redondance, ensuite la préservation de la lexicalisation, le renforcement ou le maintien de l'intonation, enfin la mise à contribution de l'interlocuteur. Il en résulte un type particulier d'énonciation pragmatique dont les variables ont été transformées dans leurs rapports d'importance respective, pour satisfaire les fins de la communication.

4. CONCLUSION

Au total, l'agrammatisme serait une forme d'économie pragmatique caractérisée surtout par une maladie de l'instanciation et de la référenciation, le contexte étant à la fois facilitateur sur le plan de la communication et inhibiteur sur le plan linguistique. Autrement dit, dans l'acte de langage agrammatique, toute opération met en jeu la performance linguistique du locuteur, la soumet aux influences de l'auditeur, aux nécessités du repérage situationnel et référentiel, à celles de l'efficience langagière, toutes composantes conditionnant le message dont la difficulté se résout dans une sorte de compromis économique faisant de l'auditeur un récepteur créatif. Mais de tous ces paramètres, c'est sans doute le repérage du message dans le contexte situationnel ou énonciatif qui semble le plus difficile à réaliser et fait de cette pathologie une maladie de l'énonciation.

BIBLIOGRAPHIE SUCCINCTE

[GOO97] Goodglass H., "Agrammatism in aphasiology", Clinical Neuroscience, 4(2) : 51-6.

[GRO95] Grodzinsky Y., "A restrictive theory of agrammatic comprehension",Brain & Language, 50(1) : 27-51.

[MAH95] Maher L.M. Chatterjee A., Gonzalez Rothi L.J., Heilman K.M., " Agrammatic sentence production : the use of a temporal-spatial strategy", Brain & Language, 49(2): 105-24.

[MCE95] McEntee L.J., Kennedy M.,"Profiling agrammatic spoken language : towards a government and binding framework", European Journal of Disorders of Communications, 3O(3): 317-32.

[NES79] Nespoulous J.L., " Approche linguistique de divers phénomènes d'agrammatisme", U. de Toulouse-Le Mirail, 1979.

[TIS73] Tissot A., Mounin G., Lhermitte F., "L'agrammatisme", Bruxelles, Dessert, 1973.

Contribution des indices métriques à l'apprentissage d'un langage artificiel

Marie-Hélène Banel[1], Uli. H. Frauenfelder[1], Pierre Perruchet[2]

[1]Université de Genève, FAPSE, Laboratoire de Psycholinguistique Expérimentale
9 route de Drize, 1227 Carouge, SUISSE
Tél.: 00 41 22 705 97 41 - Fax: 00 41 22 705 98 55
e-mail: banel@fapse.unige.ch, frauenfe@uni2a.unige.ch

[2]Université de Bourgogne, LEAD - CNRS, 6 Bd Gabriel, 2100 Dijon FRANCE
e-mail: perruche@satie.u-bourgogne.fr

ABSTRACT

This study examines the role of distributional and metrical cues in the learning of an artificial language. French subjects first heard a sample of the language (a continuous string of nonce syllables) and then were required to decide whether bi- or trisyllabic items belonged to the language or not. One group listened to a sample without metrical cues, while another group received the same language with metrical segmentation cues - alternation of short and long (accented) syllabes. The presence of these metrical cues improved the discovery and the learning of the artificial words. The implications of this research for lexical segmentation and the learning of a new language learning are discussed

1. INTRODUCTION

L'acquisition de la langue maternelle est actuellement expliquée, en partie, grâce à des mécanismes d'extraction implicite des régularités de la langue [Saf96a]. Pour apprendre les mots de la langue, les enfants doivent réussir à repérer les frontières lexicales alors que le signal de parole est le plus souvent continu. L'occurrence régulière de certaines syllabes les unes après les autres permettraient d'en inférer qu'il s'agit de la même unité (un même mot). Non seulement les enfants sont sensibles aux probabilités d'apparition successive des syllabes de la langue, mais aussi à des indices prosodiques réguliers qui indiquent des frontières de mots [Jus97]. Le problème de la segmentation des mots se pose également à des adultes qui apprennent une nouvelle langue. Aux débuts de l'apprentissage, ils perçoivent un flux continu de parole sans signification. Des recherches récentes ont étudié l'acquisition d'un langage artificiel par des adultes [Saf96b]. Il s'agissait d'un mini-langage oral très rudimentaire composé de 'mots artificiels' de format trisyllabique. Les 'mots' étaient mélangés au hasard, puis présentés les uns à la suite des autres, sans indice de frontière. Les syllabes constitutives d'un même 'mot' apparaissaient en général beaucoup plus souvent les unes à la suites autres que les syllabes

appartenant à des mots différents. Les auditeurs écoutaient le langage en ayant juste pour consigne de ''rechercher les mots qui le composaient''. Ensuite, ils devaient choisir laquelle de deux séquences de syllabes était un mot du langage artificiel. Les résultats ont montré que les participants ont utilisé les probabilités d'apparition relatives des syllabes les unes par rapport aux autres pour construire une représentation des 'mots' artificiels, et pour les mémoriser. Les indices distributionnels sont donc impliqués dans la segmentation d'une nouvelle langue.

D'autres types d'indices sont aussi utilisés lors de l'apprentissage d'une langue artificielle orale [Vro97]. Un même langage a été présenté à trois groupes d'auditeurs: des finlandais, des néerlandais et des français. La présence d'un accent distinctif initial pouvait être utilisé comme indice de frontière lexicale par les finlandais et les néerlandais car de nombreux mots commencent par une syllabe accentuée dans leurs langues. Les auditeurs finlandais ont mieux appris les 'mots' artificiels qui comportaient les caractéristiques phonologiques de leur langue: ils ont mieux mémorisé les mots accentués au début (l'accentuation était réalisée par des changements de fréquence fondamentale), ainsi que les mots présentant l'harmonie vocalique de leur langue (certaines voyelles ne peuvent apparaître dans le même mot). Les néerlandais n'ont pas utilisé l'harmonie vocalique qui n'a pas de valeur phonologique dans leur langue, mais ils ont mieux repéré et mémorisé les mots portant un accent sur la première syllabe. Les perfomances des auditeurs français ont été les plus mauvaises. Ils n'ont été sensibles ni à l'harmonie vocalique, ni à la présence d'un accent initial. L'harmonie vocalique n'a pas de valeur phonologique en français et l'accent n'a pas de valeur lexicale distinctive. Les auteurs ont conclu que les auditeurs appliquent les routines de segmentation spécifiques à leur langue maternelle, même quand ils doivent apprendre une nouvelle langue.

Les performances des français dans l'étude précédemment décrite peut s'expliquer d'après ce que nous connaissons de la contribution des indices prosodiques à la segmentation lexicale du français. L'accent en français n'a pas de valeur lexicale distinctive comme en néerlandais mais sa position est fixe, donc prévisible. Il est placé sur la syllabe finale des mots plurisyllabiques et se marque par un allongement perceptivement saillant de cette syllabe par rapport aux précédentes. Malgré la tendance à accentuer la syllabe initiale dans certains styles de parole (ex. le style radiophonique), la structure bref-long apparaît comme la structure métrique de base [Vai92]. Dans des expériences de jugement des auditeurs français devaient juger si des séquences ambigües telles que 'bordure' correspondaient à un seul ou deux mots. Les résultats ont mis en évidence que le pattern métrique long-bref inhabituel pour des séquences bisyllabiques facilite la segmentation alors que le pattern bref-long usuel pour les bisyllabes conduit à percevoir un seul mot [Ban94]. Les français utilisent une stratégie de segmentation métrique qui s'applique indépendamment du traitement de paramètres structuraux syllabiques [Ban97]. Les faibles performances des participants français dans l'étude de [Vro97] dans la condition avec indices prosodiques s'explique par le fait que l'accentuation des mots des 'mots' artificiels devait les induire à opérer des segmentations erronées du langage artificiel. De plus, la durée, et non pas les variations de fréquence fondamentale qui accompagnent l'accentuation, est l'indice majeur utilisé pour la segmentation lexicale en français [Rie80].

L'expérience que nous présentons a pour objectif de montrer dans quelles conditions d'accentuation la segmentation d'un langage artificiel et l'apprentissage de 'mots' artificiels par des francophones sont performants. La façon dont les participants mémorisent les 'mots' est également étudiée.

2. EXPÉRIENCE

L'expérience consistait à exposer à des auditeurs français à un mini-langage artificiel en leur demandant d'en rechercher les mots (phase d'apprentissage). Le langage était composé de syllabes successives: S1. S2 .S3. S4. S5. S6. S7. S8. S1. S2. S3.
Les syllabes 1.2.3 formaient un mot, les syllabes 4.5. un autre mot et les syllabes 6.7.8. encore un autre mot. Les syllabes d'un même mot (S1.S2.S3) apparaissaient 7 fois plus les unes à la suite des autres que les syllabes qui pouvaient se suivre, mais qui appartenaient à des mots différents (ex. S3. S4). Après l'écoute du langage, des essais de 2 séquences composées de syllabes de ce langage étaient présentés aux participants. Ils devaient décider laquelle des 2 était un mot de ce langage (phase test). Des mots étaient comparés à des non-mots, et des non-mots ont été comparés entre eux. 4 types de *non-mots* artificiels ont été créés pour chaque **mot** (ex de mot: **'vobdaznoul'**):

1. Une suite n'existant pas dans le langage artificiel (ex. *S2 + S7*).
2. Une suite qui recouvre des parties de 2 mots différents qui se suivent dans le langage (ex. *S3 + S4*).
3. Un début de mot (la dernière syllabe appartient à 1 autre mot) '*vobdazpuk*'.
4. Une fin de mot (la 1re syllabe appartient à 1 autre mot) '*neunedaznoul*'.

Quatre groupes de participants ont été constitués. 2 groupes expérimentaux (AMF et AMN) passaient la phase d'apprentissage et la phase test. 2 groupes contrôle (CMF et CMN) ont participé uniquement à la phase test. Le groupe AMF écoutait un langage qui présentait la métrique du français, et participait à une phase test où les séquences portaient la métrique du français. Le groupe AMN écoutait le même langage qui possédait une métrique neutre (succession de syllabes longues), et entendait des séquences portant une métrique neutre lors de la phase test. Le groupe contrôle CMF passait seulement une phase test avec métrique française. Le groupe contrôle CMN participait juste à une phase test où les séquences portaient une métrique neutre (voir Table 1).

Table 1: conditions expérimentales auxquelles ont participé les différents groupes de participants.

	Groupes de participants			
	AMF	**CMF**	**AMN**	**CMN**
apprentissage avec métrique du français	oui	non	non	non
apprentissage avec métrique neutre	non	non	oui	non
test: métrique française	oui	oui	non	non
test: métrique neutre	non	non	oui	oui

Dans l'hypothèse que les auditeurs appliquent les procédures de segmentation spécifiques à leur langue [Vro97], les résultats seront les suivants. Le groupe AMF (apprentissage avec métrique française) repérera et apprendra mieux les mots artificiels que le groupe AMN qui ne dispose pas d'indices métriques de segmentation. Les performances du groupe AMF seront supérieures à celles du groupe contrôle CMF qui n'aura pas bénéficié d'un apprentissage. Si les participants du groupe AMN (apprentissage du langage avec métrique neutre) utilisent les probabilités d'apparition de syllabes successives pour trouver les mots, leurs performances seront supérieures à celles du groupe CMN (cf. [Saf96]). L'étude des préférences entre non-mots lors de la phase test indiquera si les participants ont plutôt mémorisé les débuts ou les fins de mots. Il s'agira de savoir si, pour le groupe AMF la syllabe accentuée n'est pas mieux mémorisée que les syllabes inaccentuées. Cela s'expliquerait par la saillance perceptive de la syllabe accentuée, et par le fait que les syllabes

accentuées peuvent être utilisées pour guider l'attention des auditeurs [Pit90].

2.1 Méthode

Matériel et procédure
18 syllabes ont été choisies pour former le langage artificiel. Toutes étaient sans signification, de structure CVC (la plupart des syllabes CV sont des mots en français), et n'appartenaient pas au français. Ce choix a été motivé par la constatation que les syllabes sans signification n'apparaissent pas à la même fréquence au début et à la fin des mots français. Chaque syllabe a été produite isolément sur un ton neutre et à débit constant par un français. Nous nous sommes assurés auprès de 5 auditeurs que les syllabes sans signification étaient bien perçues comme telles et ne pouvaient être prises pour des mots français.

4 mots artificiels bisyllabiques et 4 mots trisyllabiques ont été construits à partir de ces syllabes. Seules 2 syllabes ont été utilisées dans des mots différents. Pour chaque **mot** tel que '**vobdaznoul**' 4 types de *non-mots* ont été construits à partir des 18 syllabes: 1 séquence qui n'existe pas dans le langage '*banrbonfpouk*', 1 séquence qui chevauche 2 mots '*chulpukbanf*', 1 début de mot '*vobdazpuk*' et 1 fin de mot '*neunedaznoul*'. Les deux derniers types de *non-mots* devaient permettre de s'assurer que les participants avaient bien formé une représentation des mots entiers et non une représentation fragmentaire. Chaque **mot** et *non-mot* allait être construit en 2 versions: 1 avec la métrique française et 1 avec une métrique neutre.

Une étude des production de 5 français a été réalisée pour déterminer un modèle de durée brève et de durée longue pour chaque syllabe, ainsi que le pattern métrique prototypique des bisyllabes et des trisyllabes artificiels. Une même syllabe a été placée au début, au milieu et à la fin de séquences sans signification trisyllabiques ou bisyllabiques. Les analyses ont montré que le pattern métrique des bisyllabes était bien bref-long et le pattern des trisyllabes bref-bref-long. La 1re syllabe des trisyllabes avait tendance à être plus longue que la 2de mais la différence de durée n'était pas significative. Les données confirment les patterns généraux obtenus par [Pas92] pour des mots. Les données du locuteur possédant la meilleure diction (il a produit les stimuli) ont été utilisées. Le taux moyen de compression adopté pour construire les syllabes brèves a été de 40%.

Les syllabes ont été concaténées pour former les mots (et les non-mots) artificiels. C'était nécessaire pour éviter des biais dûs aux différences de coarticulation entre syllabes de mots successifs différents. Avant concaténation, la durée de certaines syllabes a été diminuée pour pouvoir former les différents patterns métriques des stimuli. Les polysyllabes ont été fabriqués en 2 versions: une avec la métrique française (pattern des bisyllabes: bref-long, trisyllabes: bref-bref-long), et une avec la métrique neutre (pattern des bisyllabes: long-long, trisyllabes: long-long-long). L'ordre de présentation des mots artificiels utilisés pour former un langage artificiel continu a été généré automatiquement. Chaque mot artificiel apparaissait 100 fois dans le langage. Une séquence de syllabes constitutives d'un même mot apparaissait 7 fois plus qu'une séquence de syllabes de même longueur et appartenant à des mots différents. Un même mot ne pouvait se répéter immédiatement. Le même ordre de présentation a servi à construire le langage dans la version avec la métrique française et dans la version présentant la métrique neutre. Les mots du langage ont été concaténés automatiquement.

Les 4 groupes de 12 participants qui ont passé l'expérience sont présentés en Table 1. Les groupes expérimentaux (AMF, AMN) ont reçu pour consigne d'écouter le langage artificiel et d'en rechercher les mots, sans avoir d' autre information. Lors de la phase test, les participants savaient que 2 séquences de syllabes séparées par un petit silence (1s) allaient être présentées. Ils devaient choisir laquelle des 2 séquences (ex: **mot**+ *non-mot*, ou *non-mot*+ *non-mot*) était un mot du langage artificiel (l'ordre de présentation a été contrebalancé). Les groupes contrôle (CMF, CMN) ont participé à la 2de phase sans passer par l'apprentissage. Ils devaient choisir laquelle des 2 séquences pourrait être un mot qu'on utiliserait pour construire un langage artificiel. L'apprentissage durait 18 minutes, la phase test durait 15 minutes.

Participants
48 monolingues français étudiants en psychologie à l'Université de Bourgogne ont passé l'expérience.

2.2 Résultats

Le pourcentage de mots reconnus comme des mots du langage artificiel a été calculé. Des différences ont été obtenues en fonction du groupe de participants (voir Table 2). Les groupes contrôle ont répondu de façon aléatoire, alors que les groupes expérimentaux ont repéré et mémorisé des mots.

Table 2: % de mots reconnus comme des mots du langage artificiel (les groupes de sujets sont entre ( )).

	Langage avec métrique	
	du français	neutre
apprentissage	**85 (AMF)**	**67 (AMN)**
contrôle	**50 (CMF)**	**53 (CMN)**

Les analyses ont été effectuées avec le nombre de mots reconnus comme variable dépendante. Les performances du groupe AMF (apprentissage avec métrique française) ont été meilleures que celles du groupe AMN (apprentissage avec métrique neutre) ($t22 = 4.68$, $p<.0001$), et que celles du groupe contrôle CMF ($t22 = 14.11$, $p<.0001$). Les performances du groupe AMN étaient supérieures à celle de leur groupe contrôle CMN correspondant ($t\ 22 = -3.31$, $p<.002$).

Lors de la phase test les mots étaient comparés à 4 types de non-mots (essais **mot** + *non-mot*), notamment à des débuts et fins de mots (ex. le **mot 'vobdaznoul'** a été présenté avec le *non-mot* '*vobdazpuk*', et avec le *non-mot* '*neunedaznoul*'). Il s'agissait de s'assurer que les participants avaient bien formé une représentation complète des mots, et non fragmentaire. Le nombre de mots correctement reconnus comme des mots lors de toutes les présentations avec des non-mots (essais **mot**+*non-mot*) est présenté en Table 3.

Table 3: nombre moyen de mots toujours reconnus comme des mots du langage et % correspondant.

	Langage avec métrique du français		Langage avec métrique neutre	
groupes	AMF		AMN	
	4.9 (σ = 1.1)	61%	2.5 (σ = 1.8)	31%
groupes	CMF		CMN	
	0.4 (σ = 0.4)	5%	1 (σ =0.7)	12%

Le groupe AMF a mémorisé plus de mots que le groupe AMN ($t22 = 4.92$, $p<.0001$), et plus que le groupe CMF ($t\ 22 = 12.24$, $p<.0001$). Les résultats du groupe AMN diffèrent de ceux du groupe CMN ($t22 = 2.63$, $p<.02$).

Les réponses des participants aux essais où ils décidaient lequel de 2 *non-mots* correspondait le plus à un mot du langage ont été analysées. Seuls les choix entre les débuts de mots ('*vobdazpuk*') et les fins de mots ('*neunedaznoul*' pour le **mot 'vobdaznoul'**) sont exposés. Le groupe AMF a choisi les fins de mots aux débuts de mots dans 66 % des cas, le groupe AMN dans 54 % des cas. La différence entre le nombre de fins de mots et le nombre de débuts de mots n'est significative que pour le groupe AMF (apprentissage avec métrique française) (groupe AMF: $t11 = 3.189$, $p<.01$; groupe AMN: $t<1$).

3. Discussion

Selon l'hypothèse de l'application de la procédure de segmentation de sa langue maternelle lors de la perception d'une autre langue [Vro97], les participants AMF devaient avoir mieux repéré et mémorisé les mots artificiels que les participants AMN. Les résultats vont dans le sens de l'hypothèse. Les auditeurs développent différentes procédures de segmentation en fonction des propriétés phonologiques de leur langue [Vro97]. La procédure qui est pertinente pour une langue ne le sera pas pour une autre. Cette inadaptation risque de retarder l'apprentissage des mots de la nouvelle langue. Les recherches dans ce domaine restent à développer.

Par ailleurs, les résultats ont mis en évidence que le groupe AMF a préféré choisir comme mot des non-mots correspondant à des fins de mots que des non-mots étant des débuts de mots. Le groupe AMN n'a pas manifesté cette préférence. Il faut noter que les fins de mots pour le groupe AMF comportaient la syllabe accentuée des mots artificiels. Cela suggère que le groupe AMF a mieux mémorisé les syllabes accentuées car elles sont saillantes perceptivement. Ces syllabes seraient utilisées pour guider l'attention des auditeurs sur certaines parties du signal (cf. [Pit90]). La question se pose de savoir si les mots artificiels sont mémorisés avec leurs propriétés acoustiques propres, ou indépendamment de ces propriétés. Les relations entre segmentation, abstraction de mots et apprentissage restent à spécifier.

4. Remerciements

La recherche a été financée par une Bourse de la Fondation Fyssen accordée au premier auteur, et par le FNRS 1113-049698.96. N. Nguyen a prononcé les syllabes du langage, et fourni le programme de concaténation des syllabes. M. Fayol nous a accueillis pour expérimenter dans le laboratoire qu'il dirige.

5. Bibliographie

[Ban94] Banel, M-H., Bacri, N. (1994), 'On metrical patterns and lexical parsing in French', Speech Communication, 15, 115-126.

[Ban97] Banel, M-H., Bacri, N. (1997), 'Reconnaissance de la parole et indices de segmentation métriques et phonotactiques', L'Année Psychologique, Vol. 97 (1), 77-112.

[Jus97] Jusczyk, P. W. (1997), 'The discovery of spoken language', Cambridge, MA:MIT Press.

[Pas92] Pasdeloup, V. (1992), 'Durée intersyllabique dans le groupe accentuel français', Actes des 19es Journées d'Etude sur la Parole, Bruxelles.

[Pit90] Pitt, M., Samuel, A. (1990), ''The use of rhythm in attending to speech'', JEP: HPP, Vol. 16 (3), 564-573.

[Rie80] Rietveld, A. C. (1980) "Word boundaries in the French Language", Language and Speech, Vol. 23 (3), pp. 289-296.

[Saf96a] Saffran, J. R., Aslin, R. N., Newport, E.L. (1996), 'Statistical learning by 8-month old infants', Science, Vol. 274, pp. 1926-1928.

[Saf96b] Saffran, J. R., Newport, E.L, Aslin, R. N.. (1996), Word segmentation: the role of distributionnal cues, J. Of Memory and Language, Vol. 35, pp. 606-621..

[Vai92] Vaissière, J. (1992), Rhythm, accentuation and final lengthening in French, in Sundberg L., Nord L. & Carlson R. (Eds), *Music, Language, Speech and Brain*, Wenner-Gren International Symposium Series, Vol. 59, pp. 108-120.

[Vro97] Vroomen, J., Tuomainene, J. De Gelder, B. (soumis), Language specific cues to speech segmentation: the role of vowel harmony and lexical stress in Finnish, French and Dutch.

Le rôle de la syllabe : segmentation lexicale ou classification?

Nicolas Dumay[1], Marie-Hélène Banel[2], Uli H. Frauenfelder[2] & Alain Content[1,2]

[1] Laboratoire de Psychologie Expérimentale,
Université Libre de Bruxelles — 50 avenue F.D. Roosevelt — CP-191 — 1050 Bruxelles — BELGIQUE
Tél: ++322 650 42 29 — Fax: ++322 650 22 09 — e-mail: ndumay@ulb.ac.be

[2] Laboratoire de Psycholinguistique Expérimentale,
Université de Genève — SUISSE

ABSTRACT

Two word-spotting experiments were conducted to assess the role of syllable onsets and offsets in lexical segmentation. Participants detected CVC words embedded initially or finally in bisyllabic nonce strings with aligned CVC.CVC or non-aligned CV.CCVC syllabic structure. Misalignment of word and syllable onsets in the spotting of item-final words produced a greater perceptual cost than the misalignment of word and syllable offsets. These results support the view that the syllabic structure of continuous speech is mainly used to locate potential word onsets. The implications of these findings for the nature of spoken word recognition are discussed.

INTRODUCTION

La compréhension de la parole nécessite la reconnaissance des mots qui la constituent. Or, la parole est continue et les frontières de mots sont souvent masquées. L'auditeur doit donc trouver l'alignement correct entre les unités signifiantes contenues dans le signal et ses propres représentations lexicales: c'est le problème de la segmentation lexicale [pour une vue d'ensemble, Fra96]. S'il existe des corrélats acoustiques des frontières de mots, leur faible fiabilité a conduit à postuler l'implication de deux autres types d'information: les connaissances lexicales, d'une part, et les régularités dues à la structuration de la chaîne phonémique et à la métrique, d'autre part [pour une approche intégrée, Nor97].

Selon l'hypothèse lexicale, la segmentation découle de la compétition entre les multiples candidats activés par une même portion du signal. De cette compétition résulte une séquence constituée des mots les plus activés: celle-ci rend compte de l'entièreté du signal, et les mots vainqueurs ne se chevauchent à aucun moment [appuyant cette hypothèse, les résultats de Luc90; McQ94; Nor95; Vro95; pour des simulations, McC86; Fra90; Nor94].

L'autre type d'information potentielle concerne les régularités que l'on peut extraire du signal de parole afin de contraindre le processus d'accès lexical. Dans ce cadre, la structure syllabique du signal pourrait fournir des indices particulièrement importants en français [Kol95; Pal93; Seg90].

D'autre part, l'utilisation de la structure métrique de la langue est amplement démontrée. Pour les langues à réduction vocalique, comme l'anglais et le néerlandais, la Stratégie de Segmentation Métrique (MSS), postulée par Cutler et Norris, serait telle que la qualité "pleine" d'une voyelle (en opposition à "réduite") déclencherait l'accès lexical à la frontière initiale de la syllabe dont elle est le noyau [Cut88; Vro96]. En français, la systématicité de la position de l'accent métrique sur la dernière syllabe du mot permettrait, sur base des contrastes temporels, la formation de groupes rythmiques correspondant à des mots [Ban94].

Si l'hypothèse d'une stratégie métrique est correcte, elle nécessite de localiser la frontière initiale des syllabes à voyelle pleine. Dans cette perspective, les débuts de syllabes sont des points d'alignement privilégiés à partir desquels déclencher l'accès lexical. Seul le repérage des débuts de syllabe serait crucial pour la reconnaissance des mots. Une étude examinant les choix de segmentation syllabique explicite a montré une plus grande consistence dans la détermination des débuts de syllabe par rapport aux fins de syllabe [Con98].

Par contre, une autre conception du rôle de la syllabe insiste sur la classification du signal en unités syllabiques. Ainsi, le modèle d'accès lexical proposé par Segui et al. [Seg90] postule que le signal de parole est recodé en unités de taille syllabique utilisées pour accéder au lexique.

Nous avons tenté de départager ces deux conceptions, fonction de segmentation vs. fonction de classification, par une tâche de détection de mots dans des séquences de parole sans signification. Dans une première expérience, les mots monosyllabiques à détecter [e.g. *"lac"*] étaient enchassés à la fin de bisyllabes sans sens, et leur début étaient alignés [e.g. *"zun.**lac**"*] ou pas [e.g. *"zu.**glac**"*] avec le début de la seconde syllabe de la séquence[1]. Dans une seconde expérience, les mots à détecter étaient enchassés au début des bisyllabes, et cette fois, c'était leur fin qui était alignée [e.g. *"**lac**.tuf"*] ou pas [e.g. *"**la**.cluf"*] avec la fin de la première syllabe de la séquence.

Pour les enchassés en finale, les deux approches font la même prédiction: la détection devrait être facilitée lorsque début de mot et début de syllabe correspondent, par rapport à la condition de non-alignement qui nécessite une resyllabation de la séquence réassignant la première consonne du groupe médian en coda de la première syllabe. Par contre, pour les enchassés à l'initiale, elles font des prédictions divergentes. L'hypothèse de la segmentation à l'attaque ne prédit pas d'avantage pour la correspondance entre fin de mot et fin de syllabe sur la condition de non-alignement, la frontière finale de la syllabe n'étant pas pertinente pour l'accès lexical.

[1] Le symbole "." indique une frontière syllabique.

L'hypothèse de classification quant à elle prédit une détection plus aisée lorsque fin de mot et fin de syllabe sont alignées, et un coût de resyllabation équivalent à celui entraîné par le non-alignement de début.

EXPÉRIENCE I

L'expérience I visait à déterminer l'effet de l'alignement pour la frontière initiale.

Méthode

Participants. Trente-deux étudiants de l'Université de Genève ont participé à l'expérience, tous de langue maternelle française et ne présentant aucun trouble auditif ou du langage.

Matériel et procédure. Trente mots (CVC) français de contenu apparaissaient comme cible à détecter en position finale dans des séquences bisyllabiques CVCCVC sans signification. La première syllabe des séquences était elle-même un non-mot.

Deux conditions ont été étudiées. Dans la condition d'alignement, les consonnes médianes du bisyllabe appartenaient nécessairement à des syllabes différentes [C.C], de sorte que le début du mot cible était aligné avec le début de la seconde syllabe [e.g. *"zun.lac"*][2]. La séquence consonantique utilisée était impérativement attestée à l'intérieur des mots français. Dans la condition de non-alignement, les contraintes de structuration groupaient les consonnes médianes en attaque de la seconde syllabe [.CC], de sorte que le début du mot cible était décalé d'un phonème par rapport au début de la seconde syllabe [e.g. *"zu.glac"*].

Trente bisyllabes sans signification et dont aucune syllabe n'était elle-même un mot ont servi de remplisseurs: 15 séquences CVC.CVC [e.g. *"zil.ruf"*] et 15 séquences CV.CCVC [e.g. *"na.dric"*]. La distribution des séquences consonantiques dans les remplisseurs calquait celle dans les cibles. Essais-tests et remplisseurs ont été mélangés aléatoirement en deux listes expérimentales, chacune composée de 30 essais-tests et 30 remplisseurs. La position séquentielle des remplisseurs et des cibles était identique d'une liste à l'autre, le facteur alignement/non-alignement étant contre-balancé à travers les listes.

Le matériel, prononcé par un locuteur de langue maternelle française, a été stocké à une fréquence d'échantillonnage de 16 kHz avec résolution 16-bit sur le disque dur d'un ordinateur Macintosh IICI. L'expérience était gérée par le logiciel Psyscope 1.1. A chaque essai, un bip de 100 ms était présenté 1 s avant le stimulus. L'intervalle inter-stimuli était de 3 s.

La tâche du sujet était de presser le plus rapidement possible un bouton-réponse dès qu'il détectait la présence d'un mot à la fin d'une des séquences, et ensuite de répéter à voix haute le mot détecté.

[2] Les séquences consonantiques utilisées étaient attestées dans BRULEX [Con90]. La position de la frontière syllabique a été déterminée en prenant en compte le Principe d'Attaque Maximale, les règles phonotactiques, et l'échelle de sonorité.

Résultats et discussion

Le Temps de Réaction (TR) mesuré à partir de la fin de la cible et le Taux d'erreurs ont été pris comme variables dépendantes. Les TRs situés à plus de 2 écarts-types au-delà de la moyenne du sujet pour une condition donnée (5.5 % des données) ont été remplacés par cette moyenne. Le taux d'erreurs global (7 %) incluait les omissions ainsi que les essais auxquels le mot attendu n'a pas été rapporté (Table 1).

Table 1. TRs moyens et pourcentages d'erreurs en fonction du facteur Alignement dans les Expériences I (frontière initiale) et II (frontière finale).

Expériences	V.D.	Align.	Non-align.
Fr. Initiale	TRs (ms)	324	398
(Exp. 1)	Ers (%)	3.3	10.6
Fr. Finale	TRs (ms)	698	732
(Exp.2)	Ers (%)	15.6	15.6

Analyses des Temps de réaction. En moyenne, les mots cibles étaient détectés 74 ms plus vite lorsqu'ils se trouvaient alignés avec le début de la seconde syllabe (324 ms) que lorsqu'ils ne l'étaient pas (398 ms). Des ANOVAs pour mesures répétées avec l'Alignement comme facteur principal et le Sujet (*F1*) et l'Item (*F2*) comme facteurs aléatoires ont révélé que l'effet d'Alignement sur la latence de réponse était significatif ($F1(1,31) = 15.84$, $p < .001$; $F2(1,29) = 24.58$, $p < .001$). La latence de détection était donc significativement plus courte lorsque le mot cible était aligné avec le début de la seconde syllabe que lorsqu'il ne l'était pas.

Analyses des Taux d'erreurs. Le taux d'erreurs était en moyenne trois fois plus important lorsque le début de la cible n'était pas aligné avec celui de la seconde syllabe (10.6 %) que lorsqu'il l'était (3.3 %). Des ANOVAs similaires à celles effectuées sur les TRs ont révélé que cet effet d'Alignement sur les erreurs était significatif ($F1(1,31) = 16.31$, $p < .001$; $F2(1,29) = 5.40$, $p < .03$). Il était donc significativement plus difficile pour l'auditeur de détecter les mots non-alignés avec le début de la seconde syllabe de la séquence que les mots alignés avec lui.

EXPÉRIENCE II

L'expérience II visait à déterminer l'effet de l'alignement pour la frontière finale.

Méthode

Participants. Trente-deux étudiants de l'Université de Genève ont participé à l'expérience, tous de langue maternelle française et ne présentant aucun trouble auditif ou du langage.

Matériel et procédure. Trente mots (CVC) français de contenu apparaissaient comme cibles en position initiale dans des séquences bisyllabiques CVCCVC sans signification. La seconde syllabe des séquences était toujours un non-mot.

Dans la condition d'alignement, les consonnes médianes du bisyllabe appartenaient nécessairement à des syllabes

différentes [C.C], de sorte que la fin du mot cible était alignée avec la fin de la première syllabe [e.g. *"lac.tuf"*]. Dans l'autre condition, les contraintes de structuration groupaient les consonnes médianes en attaque de la seconde syllabe [.CC], de sorte que la fin du mot-cible était décalée d'un phonème par rapport à la fin de la première syllabe [e.g. *"la.cluf"*]. Pour le reste, le matériel et la procédure étaient similaires à ceux de l'expérience I.

Résultats et discussion

Le Temps de Réaction (TR) mesuré à partir de la fin de la cible et le Taux d'erreurs ont été pris comme variables dépendantes. Les TRs situés à plus de deux écarts-types au-delà de la moyenne du sujet pour une condition donnée (5 % des données) ont été remplacés par cette moyenne. Le taux d'erreurs global (15.6 %) incluait les omissions ainsi que les essais auxquels le mot attendu n'a pas été rapporté (Table 1).

Analyses des Temps de réaction. En moyenne, les mots cibles étaient détectés 34 ms plus vite lorsque leur fin était alignée avec celle de la première syllabe de la séquence (698 ms) que lorsqu'elle ne l'était pas (732 ms). Toutefois, des ANOVAs similaires à celles effectuées dans l'expérience 1 ont montré que cette tendance n'était pas statistiquement fiable ($F1(1,31) = 2.43$, $p < .5$; $F2(1,29) = 2.86, p < .5$).

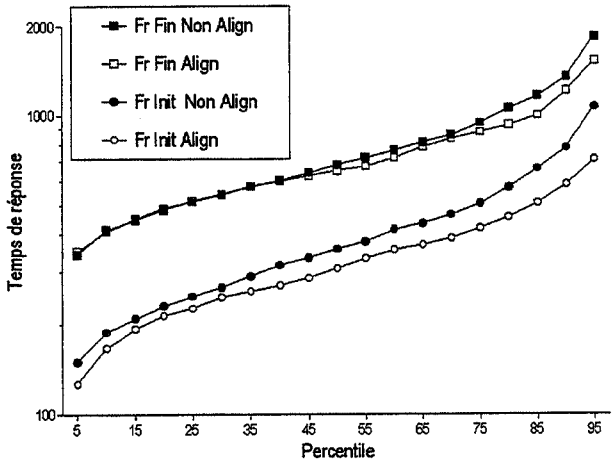


Figure 1. Distribution des TRs dans les deux expériences.

Etant donné la présence d'un petit effet de l'alignement, non significatif, dans les temps de réaction pour la frontière finale, une analyse de la distribution des TRs a été réalisée. Comme le montre la figure 1, l'effet de l'alignement n'apparaît, dans le cas de la frontière finale, que pour les latences les plus longues (à partir du percentile 75). Par contre, pour la frontière initiale, l'effet est présent sur l'ensemble de la distribution. Cette différence appuie l'hypothèse que seul l'alignement initial contraint de manière précoce et systématique le processus d'accès au lexique.

Analyses du Taux d'erreurs. Les taux d'erreurs étaient sensiblement les mêmes pour les deux conditions (alignement: 15.6 % ; non-alignement: 15.6 %). Des ANOVAs similaires à celles effectuées précédemment n'ont révélé aucune différence statistiquement significative ($Fs < 1$). La détection à l'initiale n'était donc pas plus difficile lorsque la fin du mot cible n'était pas alignée avec la fin de la première syllabe de la séquence que lorsqu'elle l'était.

DISCUSSION GÉNÉRALE

Les résultats montrent un coût perceptif de l'absence de correspondance entre frontières syllabique et lexicale dans une tâche qui implique la reconnaissance des mots. Conformément à l'hypothèse selon laquelle la structure syllabique sert principalement à déterminer les débuts de mots potentiels dans le signal, ce coût n'apparaît clairement que pour les débuts de syllabes.

Cette asymétrie est d'autant plus surprenante que les mots enchassés en finale commençaient systématiquement par /r/ ou /l/. Les sujets auraient donc pu tirer parti de la présence de cette régularité pour faciliter la localisation du début du mot. Par contre, la consonne finale des mots enchassés à l'initiale était plus variable, et ne permettait dès lors pas d'utiliser de telles stratégies. Il semble donc que l'effet négatif d'un non-alignement à l'initiale constitue un résultat robuste qui résiste à des stratégies plus spécifiques de type attentionnel.

Dans une étude analogue, Banel et Bacri [Ban97] ont observé un effet de la structure syllabique pour des monosyllabes enchassés à l'initiale. Leurs résultats semblent donc en contradiction avec ceux rapportés ici. Il faut cependant noter que dans leur étude, les séquences consonantiques utilisées dans la condition de correspondance étaient généralement illégales ou peu fréquentes en français à l'intérieur des mots [e.g. *"lamp.zok"*]. De telles séquences pourraient dès lors constituer des indices phonotactiques forts pour déterminer des frontières de mots plutot que de syllabes; ces résultats ne sont donc pas incompatibles avec notre hypothèse.

Nos résultats sont consistants avec deux études récentes en néerlandais. La première suggère qu'un mot monosyllabique enchassé en finale n'est activé que lorsqu'il est aligné avec la frontière initiale de la syllabe [Vro97]. La seconde, analogue à la présente recherche, [McQ95] a révélé une asymétrie dans l'effet de non-alignement entre mot enchassé et syllabe en fonction du type de frontière. Comme dans notre étude, le coût perceptif de la non-correspondance était plus important dans le cas de la frontière initiale [e.g. *"fim.rok" vs. "fi.drok"*] que dans celui de la frontière finale [e.g. *"vel.brul" vs. "velm.brul"*].

En conclusion, les résultats soutiennent l'hypothèse que la structure syllabique est utilisée en français pour déterminer des débuts de mots potentiels. La stratégie de segmentation métrique [Cut88], selon laquelle les attaques de syllabes à voyelle pleine déterminent des débuts de mots potentiels, pourrait donc s'appliquer non seulement aux langues à réduction vocalique (anglais et néerlandais) mais aussi en français où toutes les voyelles sont pleines. Cette similarité suggère que, contrairement à une idée répandue en psycholinguistique, la segmentation syllabique à l'attaque pourrait constituer une propriété générale des mécanismes de reconnaissance des mots indépendamment de la langue.

REMERCIEMENTS

Cette recherche a été financée par le F.N.R.S. Suisse (Projet 1113-49698.96), ainsi que par une subvention du Ministère de la Communauté française de Belgique (A.R.C. 96/01-203). M.H. Banel bénéficie d'une bourse post-doctorale de la Fondation Fyssen. N. Dumay est aspirant du F.N.R.S. (Belgique). Nous remercions Jean-Philippe Goldman pour nous avoir fourni les statistiques lexicales ainsi que Christine Meunier, Yasmine Shubber et David Ott pour leur aide dans la préparation et la passation des expériences.

RÉFÉRENCES

[Ban94] Banel, M-H. & Bacri, N. (1994). On metrical patterns and lexical parsing in French. *Speech Communication, 15*, 115-126.

[Ban97] Banel, M-H. & Bacri, N. (1997). Reconnaissance de la parole et indices de segmentation métriques et phonotactiques. *L'Année Psychologique, 97*, 77-112.

[Con98] Content, A., Kearns, R.K., & Frauenfelder, U.H. (soumis). Boundaries versus onsets in syllabic segmentation.

[Con90] Content, A., Mousty, P., & Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique, 90*, 551-556.

[Cut88] Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113-121.

[Fra90] Frauenfelder, U.H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing* (pp.50-86). Cambridge, MA: MIT Press.

[Fra96] Frauenfelder, U.H., & Content, A. (1996). La perception de la parole et la reconnaissance des mots: Apport d'une approche computationelle. *Actes des XXIèmes Journées d'Etude sur la Parole* (pp. 1-9), Avignon.

[Kol95] Kolinsky, R., Morais, J., & Cluytens, M. (1995). Intermediate representations in spoken word recognition: evidence form word illusion. *Journal of Memory and Language, 34*, 19-40.

[Luc90] Luce, P.A., Pisoni, D.B., & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing* (pp.122-147). Cambridge, MA: MIT Press.

[McC86] McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception, *Cognitive Psychology, 18*, 1-86.

[McQ94] McQueen, J.M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 621-638.

[McQ95] McQueen, J.M. & Cox, E. (1995). The use of phonotactic constraints in the segmentation of Dutch. *Proceedings of Eurospeech 95* (pp. 707-1710), Madrid.

[Nor94] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*, 189-234.

[Nor95] Norris, D., McQueen, J.M., & Cutler, A. (1995). Competition and segmentation in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1209-1228.

[Nor97] Norris, D., McQueen, J.M., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*, 191-243.

[Pal93] Pallier, C., Sebastián-Gallés, N., Felguera, T., Christophe, A., & Mehler, J. (1993). Attentional allocation within the syllabic structure of spoken words. *Journal of Memory and Language, 32*, 373-389.

[Seg90] Segui, J., Dupoux, E., & Mehler, J. (1990). The role of the syllable in speech segmentation, phoneme identification, and lexical access. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing* (pp.263-280). Cambridge, MA: MIT Press.

[Vro95] Vroomen, J. & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 98-108.

[Vro96] Vroomen, J., van Zon, M., & de Gelder, B. (1996). Cues to speech segmentation: Evidence from juncture misperceptions and word spotting. *Memory & Cognition, 24*, 744-755.

[Vro97] Vroomen, J. & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 710-720.

Rôle de la médiation phonologique dans l'apprentissage de la lecture

Ghizlaine FAUZAR & Mohamed FARID
Université Paris 7, Laboratoire de Phonétique
10, rue Charles V 75004 Paris FRANCE
Tél.: +33 01 44 78 34 59 - Fax: +33 01 44 27 79 19
E-mail: fauzar@ccr.jussieu.fr
farid@ccr.jussieu.fr

ABSTRACT

The aim of the present study was to evaluate whether Arabic-speaking children mainly use phonological mediation in the first stage of reading acquisition in a silent reading task. Thirty Moroccan children were followed from first grade to the end of second grade. We assessed their phonological skills by using a semantic decision task with phonological and visual foils (Experiment 1), and their orthographic skills by an orthographic choice task between a correct exemplar, a visual, and a phonological foil (Experiment 2). Results in the semantic decision task showed differences between visual and phonological foils. The phonological foils were more often chosen than the visual foils. In the orthographic choice task, the performance of children improved with time but errors on phonological foils were more significant. These results suggest that the phonological mediation is used by arabic speaking children in silent reading tasks.

Key words

Semantic Decision, Silent Reading, Phonological Mediation, Orthographic Representations.

1. INTRODUCTION

Les processus cognitifs régissant la lecture sont devenus le centre d'intérêt des études psycholinguistiques. Actuellement, deux types de modèles sont utilisés pour étudier le développement des habiletés de lecture. Le premier type de modèles englobe les modèles "à double voie" [Hum85], et les modèles développementaux "à étapes" [Fri86], [Mor89]. Les modèles "à double voie" présupposent l'existence d'un lexique interne dans lequel les mots sont stockés avec leurs spécifications orthographiques et phonologiques. Deux procédures permettent d'accéder aux mots en lecture. Une procédure par adressage (ou voie directe) où le mot écrit est apparié directement avec l'entrée lexicale correspondante dans le lexique orthographique interne. La procédure par voie indirecte permet de reconstituer le code phonologique des mots avant d'accéder au lexique interne, cette procédure est appelée procédure par assemblage. Le seconde classe de modèles est constituée de modèles connexionnistes qui ont essayé de simuler l'acquisition de la lecture en utilisant des réseaux connexionnistes [Sei89]. Les modèles développementaux sur lesquelles nous nous appuyons dans la présente étude sont ceux de la première classe, à savoir les modèles "à double voie" et "à étapes". Nous n'allons pas discuter de cela en détail, on précisera simplement que dans le cadre de ces modèles, les enfants sont supposés utiliser, au début de l'acquisition de la lecture, la voie indirecte. Cette procédure permet à l'enfant débutant de lire toutes les chaînes de lettres, aussi bien celles qui correspondent à des mots qu'il connaît, que celles qui renvoient à des mots qu'il ne connaît pas ou encore à des pseudomots. Au départ, cette procédure est lente et laborieuse chez l'enfant car toute son attention est dirigée sur le décodage phonologique, qu'il ne maîtrise pas encore assez, ce qui n'est pas le cas du lecteur adulte qui maîtrise les correspondances entre graphème et phonème. Tous les items, pseudomots et mots fréquents et moins fréquents sont alors traités par médiation phonologique, ce n'est que ultérieurement que se mettra en place le traitement des mots par la procédure orthographique directe.
De ce fait, les études psycholinguistiques qui ont été faites sur l'anglais [Sein 89], l'allemand [Wim90] et le français [Spr98a], [Spr98b] ont mis en évidence le fait que les enfants utilisaient essentiellement le traitement phonologique durant les premières étapes de l'acquisition de la lecture.

Les motivations principales de la présente étude, sur l'acquisition de la lecture et de l'écriture en arabe, sont expliquées par le fait qu'il n'y ait pas assez de travaux psycholinguistiques sur la langue arabe. Les autres raisons qui ont motivé cette recherche sont, plus précisément, liées aux diverses questions que l'on se pose sur les mécanismes de l'apprentissage de l'arabe au tout début de l'acquisition de l'écrit par les jeunes arabophones.

En arabe, la représentation des voyelles brèves (ou signes diacritiques) est optionnelle, surtout chez le lecteur expert, qui est capable de lire sans avoir besoin des diacritiques, par contre, il est plus difficile pour un lecteur débutant de lire des mots sans voyelles brèves. Ceci dit, la langue arabe fait partie des langues à orthographe régulière telles que l'espagnol ou l'hébreu. De ce fait la question qui nous semble pertinente est si, dans un système orthographique tel que l'arabe, les enfants arabophones utilisent la médiation phonologique au début de leur apprentissage de la lecture.

Nous avons choisi de présenter ici deux manipulations qui font partie d'un ensemble de travaux portant sur l'apprentissage de la lecture en arabe. Les deux expériences que nous présentons dans ce travail rendent compte du rôle de la médiation phonologique et de la mise en place du processus orthographique dans l'apprentissage de la lecture en arabe. La première étude est une tâche de décision sémantique qui a pour objectif d'évaluer si les jeunes lecteurs marocains utilisent la médiation phonologique au début de leur apprentissage de la lecture en arabe. La seconde étude est complémentaire de la première, elle permet quant à elle, d'examiner l'éfficience du processus orthographique.

2. Expérience 1: Décision Sémantique

Le but de cette expérience est d'évaluer comment la médiation phonologique est utilisée en lecture silencieuse. Nous avons choisi pour cela une tâche de catégorisation sémantique développée par Van Orden (1987). La tâche des enfants est de dire si les items qu'on leur présente sont des exemples d'une catégorie sémantique, par exemple si *pome* (intrus homophone) ou *pomne* (intrus visuel) est un fruit. Notre hypothèse est que si les enfants utilisaient la médiation phonologique en début de l'acquisition de la lecture de l'arabe, nous devrions trouver moins de réponses correctes pour les intrus phonologiques que pour les intrus visuels.

2.1. Méthode

Population

Nous avons choisi 30 enfants marocains tous monolingues et dont la langue maternelle est le dialecte marocain. Ils sont scolarisés dans trois écoles publiques marocaines. Ils ont été vu une première fois en Mars 1996, (leur âge moyen était de 7 ans et 2 mois), puis une deuxième fois en Avril 1997 Ces enfants n'ont aucun trouble langagier ou sensori-moteur. Leur niveau cognitif a été évalué au niveau du CP par un test d'intelligence non verbal (matrices de Raven PM 47).

Stimuli

Nous avons utilisé des catégories sémantiques accessibles aux enfants telles que la catégorie de noms de couleurs, d'animaux, de personnages, de fruits et d'objets concrets. Pour faciliter la manipulation pertinente des variables, nous n'avons utilisé que des pseudomots de deux types différents, des intrus visuels et des intrus phonologiques. Ces deux types de pseudomots ont été dérivés de mots réguliers et très fréquents existants dans la langue arabe. Par exemple, pour la catégorie des noms d'animaux, l'intrus visuel [farad¿un], et l'intrus phonologique [faras¿un] sont tous deux dérivés du mot [farasun] ("un cheval"). Les intrus visuels ne sont pas des homophones des exemples mais ils leur sont visuellement plus proche que les intrus phonologiques. Pour ce test nous précisons que les items de notre corpus ont été appariés en deux listes différentes élaborées à partir de 20 mots corrects chacune. D'un autre côté, nos sujets ont été répartis en deux groupes de 15 enfants; les enfants du 1er groupe ont été testé sur un matériel composé de 20 pseudomots cibles dont 10 intrus visuels issus de la première liste et 10 intrus phonologiques issus de la 2ème liste de telle sorte que les deux types d'intrus ne soient pas dérivés des même mots, tandis que le 2ème groupe a été testé sur un matériel composé d'un même nombre d'intrus visuels et d'intrus phonologiques mais choisis dans des listes opposés à celles du 1er groupe. A chacune des listes expérimentales, nous avons rajouté 20 distracteurs, qui sont tous des exemples corrects des catégories utilisées pour cette tâche. Nous précisons que les mots corrects à partir desquels ont été dérivés les pseudomots de l'expérience ne figuraient pas dans la liste expérimentale.

Tâche et procédure

Le test a été administré à chaque enfant individuellement sans utilisation d'ordinateur. On présente en premier le nom d'une catégorie oralement à l'enfant et immédiatement après, l'item cible (distracteur, ou intrus visuel ou intrus phonologique) lui est présenté par écrit. La tâche de l'enfant est de juger si l'item présenté est un exemple de la catégorie sémantique qui vient d'être précisée. Par exemple, l'expérimentateur pose la question suivante "est-ce un animal?" et juste après l'item cible apparaît. Les enfants devaient répondre oralement par OUI ou NON. Le score de chaque enfant était reporté sur une fiche. Les réponses correctes sont représentées par "1" et les erreurs par "0"

2.2. Résultats et discussion

Les résultats concernant la moyenne des réponses correctes dans la tâche de décision sémantique sont présentés dans le tableau 1. Dans la mesure où nous n'avons pas trouvé de différence significative entre nos deux groupes de sujets, nos analyses ont porté sur l'ensemble du groupe.
Nous avons effectué une analyse de variance (ANOVA) sur les moyennes des réponses correctes pour voir si le facteur Type d'intrus (visuel vs phonologique) et le facteur Session (session 1 et session 2) ont des effets significatifs sur les performances des enfants. Ces tests ont montré que les deux facteurs ont des effets significatifs (Session, $F(1,29)= 15.04$, $p<0.01$; Type d'intrus, $F(1,29)= 31.76$, $p<0.01$). Nous observons donc que les performances.des enfants augmentent entre les sessions et les intrus phonologiques sont moins bien détéctés que les intrus visuels Cependant, l'interaction entre ces deux facteurs n'est pas significative ($F(1,29)= 1.95$, $p>0.10$). Toutefois, des analyses complémentaires indiquent que l'effet de la session est significatif pour les intrus visuels (le nombre de réponses correctes augmente significativement entre les deux sessions $F(1,29)=13,50$; $P<0,01$), mais pas pour les intrus phonologiques $F(1,29)=2,86$ NS. Il y a donc une amélioration des performances entre les sessions pour les intrus visuels mais pas pour les intrus phonologiques.

Tableau 1: La moyenne des réponses correctes en décision sémantique (les écarts-type sont mis entre parenthèses).

	Intrus Visuel	Intrus Phonologique
Session 1	3.50 (2.18)	2.13 (1.78)
Session 2	5.20 (1.24)	2.93 (1.74)

On peut dire que les données de la présente expérience confirment notre hypothèse de départ, les intrus phonologiques sont moins bien repérés que les visuels.. En d'autres termes, ces résultats indiquent que les jeunes enfants apprenant la lecture de l'arabe ont recours à la médiation phonologique en lecture silencieuse dans une tâche de décision sémantique. De plus, nos résultats indiquent qu'il n'y a pas de progression pour les intrus phonologiques entre les sessions dans leur performance, alors qu'on observe une amélioration assez nette pour la détection des intrus visuels. Ces résultats témoignent du poids de la phonologie au cours de l'apprentissage de la lecture en arabe, pas simplement en tout début de l'apprentissage (en Mars de la 1ère année du primaire) mais également dans les étapes ultérieures de l'apprentissage (en juin de la 2ème année du primaire). Cependant on peut se demander si ces résultats ne sont pas dus à l'imprécision des connaissances orthographiques de nos sujets. L'expérience suivante de vérification orthographique a pour objectif d'évaluer cette hypothèse.

3. Expérience 2: Vérification Orthographique

Le but de cette expérience est d'évaluer les aptitudes orthographiques des enfants et de voir si leur connaissance de l'orthographe des mots présentés dans l'expérience précédente n'influençait pas les résultats. Nos hypothèses sont les suivantes:

Hypothèse 1: on s'attend à ce que les performances des enfants augmentent entre les deux sessions.

Hypothèse 2: Si les enfants ont surtout recours à la médiation phonologique, on s'attend à observer plus d'erreurs sur les intrus phonologiques que sur les intrus visuels.

3.1. Méthode

Population

Il s'agit du même groupe d'enfants testés dans l'expérience précédente.

Stimuli

Pour cette tâche de vérification orthographique, nous avons présenté 10 séries de trois items comportant chacune le mot correct correspondant aux deux intrus utilisés dans la tâche de décision sémantique, ainsi que l'intrus phonologique et l'intrus visuel. exemple: [farasun] ("un cheval"), [faras¿un] et [farad¿un].

Tâche et procédure

La tâche adoptée ici est celle du choix orthographique. Nous avons présentés aux enfants sur une même planche le mot-cible correct et ses deux intrus visuel et phonologique. La tâche de l'enfant est de choisir parmi ces trois items le "bon" mot, c'est-à-dire le mot qui est orthographiquement correct. Cette tâche comme la précédente implique une lecture silencieuse.

3.2. Résultats et discussion

Les analyses statistiques de cette expérience sont représentées dans les tableaux 2 et 3 Les analyses de variances effectuées sur les réponses correctes révèlent que le facteur "Session" est significatif $F(1,29)= 19.18$, $p<0.01$. Conformément au hypothèses, les performances des enfants augmentent entre les deux sessions et leurs compétences orthographiques sont en fin de la 2ème année du primaire de l'ordre de 75% de réponses correctes, alors que nous observons moins de 30% de réponses correctes dans l'épreuve précédante pour les intrus phonologiques.

Tableau 2: La moyenne des réponses correctes en vérification orthographique.

	Session 1	Session 2
Moyenne	10.70	14.33
Ecart-type	3.83	3.37

En ce qui concerne les erreurs, nous avons calculé pour chaque enfant le pourcentage moyen d'erreurs pour les intrus visuels et phonologiques par rapport à l'ensemble des erreurs. On note que les enfants font plus d'erreurs sur les intrus phonologiques. Toutefois les analyses du test "t" effectuées sur les erreurs pour chaque session ont montré que seule au niveau du CP (1ère session) les intrus phonologiques sont moins bien détectés que les visuels. En effet, nous avons trouvé une différence significative entre intrus visuels et intrus phonologiques en 1ère session $t(29)=3,26$ $p<0,01$ mais pas pour la 2ème session $t(29)>1$ NS.

Tableau 3: Le pourcentage moyen des erreurs sur les intrus visuels et phonologiques en vérification orthographique (les écarts-type sont mis entre parenthèses).

	Intrus Visuel	Intrus Phonologique
Session 1	39.83 (17.04)	60.17 (17.04)
Session 2	48.39 (15.85)	51.61 (15.85)

La bonne performance des enfants en fin de 2ème année du primaire (75% des réponses correctes) dans cette tâche indique que leurs représentations orthographiques de l'arabe sont bien établies. On ne peut donc pas attribuer les résultats observés en décision sémantique en fin de 2ème année du primaire à l'imprécision des connaissances orthographiques des enfants. L'analyse des erreurs montrent également que les enfants utilisent la médiation phonologique surtout au début de l'apprentissage de la lecture (cf. les résultats de la 1ére session).

4. Conclusion générale

Les résultats observés dans cette étude montrent bien que les enfants au début de leur apprentissage de la lecture en arabe utilisent la médiation phonologique. Ainsi, dans la tâche de décision sémantique, ils ont fait plus d'erreurs sur les intrus phonologiques que sur les intrus visuels, et leurs performances ne s'améliorent pas de façon identique

entre les sessions, car seule une amélioration significative pour les bonnes réponses a été observée pour les intrus visuels. Donc, les enfants s'appuient bien sur l'information phonologique pour opérer un traitement sémantique.
D'autres part, les données obtenues en décision sémantique ne sont pas dues à l'imprécision des connaissances orthographiques de notre population, cela a été confirmé par la tâche de vérification orthographique dont les résultats ont révélé que les enfants présentent un bon niveau en orthographe ainsi qu'ils mettent en evidence le poids de la médiation phonologique dans cette tâche (cf. analyse des erreurs).

Bibliographie

[Hum85] Humphreys, G.W. , & Evett, L.J. (1985), "Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading", Behavioral and Brain Sciences, Vol. 8, pp. 689-740.

[Fri86] Frith, U. (1986), "A developmental framework for developmental dyslexia", Annals of Dyslexia, Vol. 36, pp. 69-81.

[Mor89] Morton, J. (1989), "An information-processing account of reading acquisition", In A.M. Galaburda (Ed.), From Reading to neurons (pp. 43-66). Cambridge, MA: Bradford Book, MIT Press.

[Rav47] Raven, J.C (1947, rééd. 1981) Progressives Matrices Standard, Issy les Moulineaux, Éditions de Psychologie Appliquée.

[Sei89] Seidenberg, M.S. & McClelland, J.L. (1989), "A distributed developmental model of word recognition and naming", Psychological Review, Vol. 96, pp. 523-568.

[Spr96] Sprenger-Charolles, L. , & Casalis, S. (1996), Lecture et écriture: acquisition et troubles du développement. Paris, PUF .

[Spr98a] Sprenger-Charolles, L. , Siegel, L.S. , & Béchennec, D. (1998), "Phonological mediation and semantic and orthographic factors in silent reading in french", Scientif Studies of Reading, 2(1), pp 3-29

[Spr98b] Sprenger-Charolles, L. , Siegel, L.S. , & Bonnet, P. (1998), "Reading and spelling acquisition in french: The role of phonological médiation and orthographic factors", Journal of Experimental Child Psychology, Vol. 68, pp 134-165

[Van87] Van Orden, G.C. (1987), A rows is a rose: Spelling, sound and reading", Memory and Cognition, Vol. 15, pp. 181-198.

[Wim90] Wimner, H. & Hummer, P. (1990) How german speaking first graders read and spell: Doubts on the importance of the logographic stage, Applied Psycholinguistics, Vol. 11, pp.349-368.

Remerciements

Nous tenons à remercier M. Georges Boulakia et Madame Liliane Sprenger-Charolles pour leurs suggestions utiles pendant la réalisation de ce travail.

ANNEXE

Exemples d'items utilisés dans les deux tâches de lecture silencieuse.

Catégorie "animal"

Mot	Intrus visuel	Intrus phonologique
سَمَكٌ	شَمَكٌ	صَمَكُنْ
samakun	$amakun	Ṣamakun
"poisson"		

Catégorie "fruit"

Mot	Intrus visuel	Intrus phonologique
تُفَّاحٌ	تُفَّاجٌ	تُفَّاحُنْ
tuffa:Hun	tuffa:jun	tuffa:Hun
"pomme"		

Segmentation en mots et compétitions lexicales

Frédéric Isel [1] et Nicole Bacri [2]

[1] Max-Planck Institute of Cognitive Neuroscience
Inselstrasse 22-26, D-04103 Leipzig
Tél.: + 49 / (0) 341 / 9940-153 - Fax: + 49 / (0) 341 / 9940 / 113 - e-mail: isel@cns.mpg.de
[2] Laboratoire de Psychologie Expérimentale, URA 316 - Université René Descartes
28 rue Serpente, F - 75006 Paris
Tél.: + 33 / 01 40 51 98 41 - Fax: + 33 / 01 40 51 70 85 - e-mail: bacri@idf.ext.jussieu.fr

Abstract

Two cross-modal priming experiments investigated whether the representation of either an initially- or a finally-embedded word may be activated when the longer carrier word is auditorily presented. Visual probes occurred at the acoustic offset of auditory primes. They were semantically related either to the embedded word or to the carrier word, or they were unrelated to the primes. In initial position the embedded word did not activate its semantic associate. In final position, there were significant priming effects for semantic associates of both the embedded word and the carrier word. It is concluded that lexical hypotheses are generated at several points during spoken word processing. Initial alignment is not crucial for lexical access. A given auditory sequence may offer the opportunity for several segmentation patterns.

1. Introduction

La reconnaissance de la parole continue implique une segmentation du signal de parole en séquences susceptibles d'être appariées avec des représentations lexicales. Nous nous proposons d'étudier deux des questions que soulève cet appariement entre entrées sensorielles et représentations mentales. Si on admet, comme le font les modèles actuels d'accès au lexique - Cohort [Mar87] et TRACE [McC86]- que l'auditeur engendre de multiples hypothèses lexicales en fonction de l'information sensorielle qu'il reçoit, il s'agit tout d'abord de déterminer si c'est en un seul point du signal (c.à.d. son début acoustique) que peut être réalisé un alignement pertinent des entrées lexicales sur cette information comme le propose Cohort, ou bien si les hypothèses sont générées continuement, selon les prédictions de TRACE. Dans le premier cas, le traitement de la parole serait fondamentalement séquentiel; dans le second cas, la reconnaissance des mots résulterait de l'issue des compétitions entre représentations activées différentiellement. Cette question se pose avec acuité quand on a à traiter des séquences de syllabes qui peuvent se concaténer en mots plus longs ou bien ne pas être recombinées, si l'une d'elles a statut de mot monosyllabique [Gow95].

La seconde question porte sur le fonctionnement des compétitions entre hypothèses lexicales au cours de la sélection, parmi les nombreux candidats, de celui qui donnera effectivement lieu à un accès [Fra96].

Le paradigme d'amorçage sémantique permet, dans sa version inter-modale, de suivre la dynamique des hypothèses lexicales au cours de l'audition des mots. Amorçes auditives et cibles visuelles sont ou non sémantiquement reliées, selon la condition. La tâche du sujet étant de décider si l'item cible est ou non un mot, la comparaison entre les temps de réponse (TR) selon la condition (reliée ou non reliée) permet d'évaluer si les propriétés sémantiques de l'amorce ont été traitées et si son activation s'est maintenue jusqu'à la réponse. L'effet d'amorçage, souvent étudié, est robuste et n'est pas affecté par la modalité sensorielle d'entrée, auditive ou visuelle [Slo94]. Les données obtenues sur le traitement de mots enchâssés avec ce paradigme permettront d'apprécier l'importance du point de départ de l'alignement entre informations acoustiques et représentations lexicales ainsi que le rôle dans la perception de la parole des appariements partiels entre signal et représentations.

On sait, par exemple, que l'audition de "honing" (miel) facilite la décision sur la cible visuelle BIJ (abeille), mais aussi que cette décision n'est pas facilitée par les amorces "woning" (appartement) ou "foning" (pseudo-mot) en dépit du fort recouvrement entre "honing" et les deux autres amorces [Mar89]. Ces données suggèrent que le traitement des séquences auditives exige un alignement initial, et le recouvrement phonologique ne semble pas susciter de compétitions entre hypothèses lexicales, qui se seraient marquées par une inhibition - un ralentissement - de la décision sur la cible BIJ. Toutefois, le modèle TRACE propose que des hypothèses multiples soient générées en différents points du signal. Le mot le plus long est certes plus activé que ses compétiteurs plus courts, mais les simulations informatiques [Fra90] et certaines données expérimentales [Vro97] montrent que, dans certaines conditions, les mots enchâssés au début d'un pseudo-mot plus long peuvent être activés transitoirement. Nous faisons l'hypothèse que les mots enchâssés sont effectivement traités, indépendamment de leur position, avant d'être inhibés par le mot plus long, mais que la possibilité de mettre en évidence leur

activation dépend de leur position dans la séquence. Transitoire, leur activation ne sera plus opérationnelle s'ils se situent au début du mot plus long ("car" dans "cargo"). Mais elle pourra se propager à la cible associée s'ils se situent à la fin de ce mot plus long ("car" dans "brancard"). Shillcock [Shi90] a ainsi montré une facilitation du traitement de la cible visuelle RIB quand elle était précédée de l'amorce auditive "trombone". L'objectif des deux expériences présentées est de montrer 1) l'effet de la position du mot enchâssé sur son activation sémantique 2) la réalité du traitement sémantique d'un mot enchâssé en position finale.

2. Expérience 1: bisyllabes contenant un mot enchâssé en position initiale

2. 1 Méthode

Sujets: Quarante sujets de langue maternelle française (français parisien)

Matériel: le matériel comporte
- 36 mots bisyllabiques monomorphémiques (durée moyenne: 520 ms) dont la première syllabe constitue un mot monosyllabique (durée moyenne: 203 ms) ont servi d'amorces auditives.
- une expérience préliminaire (20 sujets différents de ceux qui ont passé cette expérience) a permis d'associer aux 36 mots amorces 72 mots cibles qui sont présentés visuellement. Les cibles sont sémantiquement reliées aux amorces bisyllabiques ou bien au monosyllabe enchâssé que chacune des amorces contient.
- ces mêmes mots cibles étaient, dans la condition contrôle, précédés d'une amorce auditive non reliée.
- 108 paires de remplissage (36 paires mot/mot, 72 paires mot/pseudo-mot) complétaient le matériel.

Tous les mots étaient de fréquence d'occurrence moyenne ou faible. Les deux types de cibles comportaient en moyenne le même nombre de lettres. Un exemple du matériel est présenté Tableau I. Le matériel a été enregistré par un locuteur avec un débit constant et une intonation monotone.

Tableau I
Expérience 1: exemple de paires reliées et non reliées

Cible visuelle	Amorce auditive (Reliée)	Amorce auditive (Non reliée)
bateau autobus	car<u>go</u>	sulfate

Procédure: Quatre listes ont été formées de sorte que chaque sujet ne traitait une amorce ou une cible donnée qu'une seule fois. Pour chaque sujet, et afin d'éviter la mise en place de stratégies de réponse, le pourcentage d'items reliés était de 12.5 %. Chaque essai était constitué de l'audition d'une amorce suivie immédiatement de la présentation d'une cible visuelle. La tâche était de décider le plus rapidement possible si la cible était ou non un mot, en appuyant sur l'un des deux boutons de réponse. Les temps de décision lexicale étaient mesurés à partir de l'apparition de la cible visuelle.

2. 2 Résultats et discussion

Les cibles ont été traitées plus rapidement lorsqu'elles étaient sémantiquement reliées à l'amorce (519 ms, sd = 28 ms) que dans le cas contraire (537 ms, sd = 30 ms) (F1 (1,39) = 17.4, $p < .0001$), mais les cibles reliées aux monosyllabes enchâssés dans les amorces ont été traitées plus lentement que les cibles reliées aux amorces bisyllabiques (F1 (1,39) = 11.6, $p < .001$) (Tableau II). L'interaction entre la condition d'amorçage et le type de cible est significative (F1 (1,39) = 21.3, $p < .0001$). On constate ainsi que seules les amorces complètes suscitent une effet d'amorçage significatif (F1 (1,39) = 36, $p < .0001$). Les analyses de variance faites en prenant les items comme facteur aléatoire donnent des résultats analogues.

Tableau II
Expérience 1: temps de réponse moyens en ms et écarts-types corrigés pour les cibles associées aux amorces mono- et bisyllabiques dans les conditions reliée et non reliée

	Cibles reliées aux monosyllabes	Cibles reliées aux bisyllabes
Cond. reliée	536 (35)	502 (30)
Cond. non reliée	537 (37)	538 (36)
Effets d'amorçage	**+ 1**	**+ 36**

Ces résultats montrent clairement l'absence d'une activation résiduelle des formes sémantiques des mots enchâssés au début d'un mot plus long ("car" dans "cargo"). Ils confirment l'hypothèse, formulée par le modèle TRACE, que l'appariement maximum entre un signal et une représentation lexicale l'emporte, lors de la décision que prend le sujet, sur un ajustement partiel. L'alignement initial des deux mots possibles ("car" et "cargo") sur le début de la production vocale ne suffit pas pour que tous deux soient activés de façon équivalente. Toutefois, cette expérience ne permet pas de décider si le mot enchâssé ("car") a été traité et sa représentation activée, mais de façon trop transitoire pour que cette activation soit encore présente lors de l'apparition de la cible, ou bien s'il n'a donné lieu à aucun traitement. C'est sur ce point critique que porte l'Expérience 2.

3. Expérience 2: bisyllabes contenant un mot enchâssé en position finale

Si le mot enchâssé est effectivement traité, le fait qu'il soit suivi immédiatement de la cible qui lui est associée sémantiquement devrait permettre de mettre en évidence la propagation de l'activation de la représentation de l'amorce à celle de la cible qui lui est associée. L'absence d'effet significatif d'amorçage dans cette condition permettrait au contraire de conclure que seul un appariement complet entre informations acoustiques et

représentation lexicale permet d'avoir accès à un mot. Le caractère absolu de "l'avantage du mot le plus long" serait alors confirmé: ce mot ne pourrait être segmenté en ses composantes.

3. 1 Méthode

La méthode est identique à celle de l'expérience précédente, à la différence près que le matériel est constitué de 24 mots bisyllabiques monomorphémiques (durée moyenne: 515 ms) qui contiennent un mot enchâssé en position finale (durée moyenne: 215 ms). Dans la mesure du possible, les monosyllabes enchâssés sont les mêmes mots que dans l'Expérience 1 (Tableau III). Leur position en fin de séquence n'a entrainé qu'un faible allongement en comparaison avec leur durée en début de séquence. Ces 24 mots sont utilisés comme amorces. Ils sont associés à 48 cibles qui, dans la condition contrôle, sont précédées par 48 amorces non reliées. Le matériel est complété par 72 paires de remplissage (24 paires mot/mot, 48 paires mot/pseudo-mot). Pour chacun des 40 sujets, comme dans l'expérience précédente, le pourcentage de paires d'items reliés est de 12.5 %.

Tableau III
Expérience 2: exemple de paires reliées et non reliées

Cible visuelle	Amorce auditive (Reliée)	Amorce auditive (Non reliée)
autobus hôpital	brancard	sulfate

3.2 Résultats et discussion

Les cibles sont traitées plus rapidement dans la condition reliée (524 ms, sd = 28 ms) que dans la condition non reliée (562 ms, sd = 39 ms) (F1 (1,39) = 43, $p < .0001$). Les temps de réponse pour les cibles associées aux monosyllabes enchâssées dans les amorces ne sont pas significativement plus longs que les TR pour les cibles associées aux bisyllabes. Il s'ensuit - et c'est le point important- que monosyllabe enchâssé et mot porteur suscitent tous deux des effets d'amorçage significatifs (F (1,39) = 15, $p < .0003$ et F (1,39) = 38.4, $p < .0001$, respectivement), bien que l'effet associé aux bisyllabes (48 ms) soit significativement plus fort que l'effet associé aux monosyllabes (27 ms) (F (1,39) = 4.5, $p < .04$) (Tableau IV). Les résultats des analyses par item sont analogues. On notera que, bien que plus longs dans l'Expérience 2 que dans l'Expérience 1, les temps de réponse obtenus dans la condition contrôle pour laquelle les deux sous-ensembles de données peuvent être comparés- c'est à dire les TR aux cibles correspondant aux monosyllabes - ne différent pas significativement entre eux ($p>.10$).

Tableau IV
Expérience 2: temps de réponse moyens en ms et écarts-types corrigés pour les cibles associées aux amorces mono- et bisyllabiques dans les conditions reliée et non reliée

	Cibles reliées aux monosyllabes	Cibles reliées aux bisyllabes
Cond. reliée	534 (35)	515 (29)
Cond. non reliée	561 (47)	563 (47)
Effets d'amorçage	**+ 27**	**+ 48**

Ces résultats montrent que le mot court situé à la fin d'un mot plus long est effectivement traité, et que la représentation sémantique correspondante est activée et se propage à la cible associée dont elle facilite le traitement. La possibilité de concaténer les deux syllabes successives en un seul mot n'oblitère pas le traitement du mot enchâssé (cf. [Gow95]), même si l'activation sémantique générée par le mot le plus long ("brancard") est significativement supérieure à celle qui se propage du mot enchâssé ("car") à la cible qui lui est associée.

4. Discussion générale

La comparaison entre les résultats des deux expériences montre que la position du mot enchâssé dans la séquence a un effet déterminant sur la mise en évidence de son traitement. Il est clair que les formes sémantiques des mots situés à la fin d'un mot plus long sont activées, bien que leur activation soit moindre que celle du mot long qui les contient. L'activation des mots enchâssés en position finale se propage à leurs associés, et sa durée est suffisante pour faciliter la réponse des sujets. Cet effet est robuste, comme l'avait déjà montré Shillcock [Shi90]. Cette activation n'est toutefois que transitoire, comme le suggère l'absence d'un effet d'amorçage significatif lorsque le mot court est situé au début d'un mot plus long. Ce constat remet en question la généralité du concept d'alignement initial proposé par le modèle séquentiel Cohorte [Mar87]. Un mot peut être activé même s'il commence en un autre point que le début du signal de parole. Par ailleurs, nos deux expériences confirment "l'avantage du mot le plus long" tout en le relativisant. L'hypothèse d'un alignement multiple des représentations lexicales sur le signal est validée (cf. [McQ95]. Sur ce point, nos résultats appuient les propositions du modèle connexioniste TRACE.

On peut en conclure que des hypothèses lexicales sont générées dès qu'une séquence de phonèmes, de format syllabique, est présente dans le signal. L'absence de facilitation suscitée par le mot enchâssé en position initiale peut aisément s'expliquer dans le même cadre théorique. La forme lexicale de ce mot est probablement activée, mais, dès le phonème subséquent à la forme acoustique du monosyllabe, le niveau d'activation du mot porteur tend à être suffisant pour inhiber complètement la représentation du mot court. A la fin de la séquence, le mot enchâssé au début du signal est désactivé. *A contrario*, l'influence inhibitrice du mot porteur n'a pas le

temps de se propager aux représentations de ses composantes quand l'enchâssé est en position finale.

L'efficacité des inhibitions suscitées par le mot le plus long varie ainsi en fonction de la position du mot enchâssé. La recombinaison de syllabes successives en un seul mot semble ne pas être immédiate. Pendant un très bref laps de temps "brancard" et "car" sont tous deux activés. L'accès au lexique passerait par des compétitions entre candidats rivaux. Dans certains cas (Expérience 2), le système de traitement semble pouvoir contenir dans l'espace de décision simultanément deux mots dont le niveau d'activation est juste inférieur au seuil de reconnaissance. Ce cas de figure n'est pas prévu par les modèles actuels de reconnaissance des mots. Nos données montrent la pertinence d'une différenciation entre deux seuils: un seuil à partir duquel la quantité d'information acoustico-phonétique accumulée permet aux formes lexicales d'être activées, sans qu'il y ait une reconnaissance proprement dite du mot, et, d'autre part, un seuil à partir duquel les formes sémantiques associées à un mot seront activées par propagation d'un montant suffisant d'activation. C'est alors que la sélection entre formes lexicales en compétition s'opère. Cette sélection déclenche l'accès au lexique et va commander la décision. Elle est fonction de l'état du système de traitement à la fin acoustique du signal.

L'avantage du mot le plus long n'est complet que lorsque c'est sa première syllabe qui constitue un mot potentiel. Il n'est que relatif quand c'est sa dernière syllabe qui a valeur de mot. L'efficacité du mécanisme de recombinaison automatique lors de la reconnaissance des mots peut ainsi être prise en défaut. Une segmentation non appropriée est dans bien des cas possible. Un modèle de traitement qui se veut général doit intégrer cette éventualité. Eventualité d'autant plus plausible en français que la syllabe finale d'un mot plurisyllabique est généralement allongée. Si un mot potentiel est enchâssé en cette position, il bénéficiera d'un surplus d'activation, dû à ses propriétés acoustiques, qui en fait un bon candidat à la reconnaissance. A la suite de travaux antérieurs [Ban94], nous avons proposé qu'une routine spécifique de segmentation, à base métrique et prosodique, intervienne pour moduler le poids des différentes hypothèses lexicales et éviter des segmentations dont l'issue ne pourrait être intégrée à l'énoncé en cours de traitement. Notre étude de la reconnaissance des mots enchâssés suggère qu'un modèle de perception et de reconnaissance de la parole ne peut faire l'économie de routines complémentaires de segmentation pré-lexicales.

Bibliographie

[Ban94] Banel M.H. & Bacri N. (1994), "On metrical patterns and lexical parsing in French", Speech Communication, 15, pp. 115-126.

Fra90] Frauenfelder U. & Peeters G. (1990), "On lexical segmentation in TRACE: An exercise in simulation", in G. Altmann (Ed.), Cognitive models of speech processing, MIT Press, pp. 50-86.

[Fra96] Frauenfelder U. (1996), "La perception de la parole et la reconnaissance des mots: l'apport d'une approche computationnelle", XXIes JEP, pp. 1-9.

[Gow95] Gow D. & Gordon P. (1995), "Lexical and prelexical influences on word segmentation: Evidence from priming", Journal of Experimental Psychology: Human Perception and Performance, 21, pp. 344-359.

[Mar87] Marslen-Wilson W. (1987), "Functional parallelism in spoken word recognition", Cognition, 25, pp. 71-102.

[Mar89] Marslen-Wilson W. & Zwitserlood P. (1989), "Accessing spoken words: The importance of word onsets", Journal of Experimental Psychology: Human Perception and Performance, 15, pp. 576-585.

[McC86] McClelland J. & Elman J. (1986), "The TRACE model of speech perception", Cognitive Psychology, 18, pp. 1-86.

[McQ95] McQueen J., Cutler A., Briscoe T. & Norris D. (1995), "Models of continuous speech recognition and the contents of the vocabulary", Language and Cognitive Processes, 10, pp. 309-331.

[Shi90] Shillcock R. (1990), "Lexical hypotheses in continuous speech", in G. Altmann (Ed.), Cognitive models of speech processing, MIT Press, pp. 24-49.

[Slo94] Slowiaczek L. (1994), "Semantic priming in a single-word shadowing task", American Journal of Psychology, 107, pp. 245-260.

[Vro97] Vroomen J. & de Gelder B. (1997), "Metrical segmentation and lexical inhibition in spoken word recognition", J. of Experimental Psychology: Human perception and Performance, 23, pp. 98-108.

Sensibilité à la contingence phonémique et à la fréquence des unités infra-syllabiques dans les jugements phonotactiques

Ronald Peereman, Emmanuelle Ferreri, et Pierre Perruchet

L.E.A.D.- C.N.R.S., Université de Bourgogne
6 bd Gabriel, 21000 Dijon, France
Tél.: ++33 03 80 39 57 56 - Fax: ++33 03 80 39 57 67
e-mail: peereman@u-bourgogne.fr

ABSTRACT

Several recent psycholinguistic studies indicate that language acquisition is accompanied by the learning of the statistical properties of the language system. The present research investigates whether third-grade children are sensitive to phonemic contingency within the syllable unit. Vowel-coda contingency of CVC syllables was examined in Experiment 1 and onset-vowel contingency was examined in Experiment 2. In both experiments phonotactic judgements were influenced by phonemic contingency, although the effect was function of the frequency of the di-phonemic unit (vowel+coda or onset+vowel). We conclude that phonemic contingency is part of the phonotactic knowledge of 9 years old children.

1. INTRODUCTION

Ces dernières années, plusieurs recherches en psycholinguistique ont souligné le rôle des régularités statistiques entre les multiples aspects du langage dans son acquisition et son utilisation [Sei97]. Cette perspective probabilistique marque un rapprochement entre l'étude du traitement du langage et les travaux relatifs à l'apprentissage associatif [Sha95]. Elle bénéficie par ailleurs du développement des ressources informatiques fournissant des descriptions statistiques à différents niveaux d'analyse du langage [Fra96].

Plusieurs études indiquent une sensibilité précoce aux covariations statistiques dans la perception du langage. Ainsi, l'individualisation des mots dans le flux continu de la parole est du moins partiellement fonction des probabilités transitionnelles entre les syllabes [Saf96]. Dès le plus jeune âge, les enfants démontrent également une préférence pour les séquences qui respectent les règles phonotactiques de leur langue [Jus97]. L'importance de l'unité syllabique chez les jeunes enfants a toutefois conduit à une focalisation de la recherche sur la sensibilité à la fréquence et à la légalité des patrons syllabiques, peu de travaux examinant la sensibilité aux covariations phonémiques intra-syllabiques.

Diverses recherches portant sur l'apprentissage associatif indiquent un apprentissage progressif de la contingence d'événements [Sha95]. La notion de contingence entre deux événements A et B (par exemple entre deux phonèmes) peut se définir comme correspondant à la différence entre la probabilité que B (A) soit précédé (suivit) de A (B) et la probabilité que B (A) ne soit pas précédé (suivit) de A (B)[Per81]. Le but de la présente étude est d'examiner si les contingences phonémiques intra-syllabiques font partie des connaissances phonotactiques d'enfants de troisième année primaire.

Treiman et coll. [Tre98] ont récemment réalisé une tâche de jugement phonotactique consistant à juger laquelle de deux syllabes sans signification (logatomes) paraissait la plus similaire aux mots de la langue anglaise. Les données suggèrent que les préférences des enfants de 2e et 4e année primaire étaient guidées par la contingence des unités phonologiques composant la rime, c'est-à-dire la voyelle et la consonne suivante (coda). En outre, cette sensibilité à la contingence paraît augmenter avec l'âge. Cette étude ne permet toutefois pas de dissocier le rôle joué par la force de contingence des voyelles et des codas, du rôle joué par la fréquence de l'unité composite correspondant à la rime. En effet la contingence élevée de deux événements est généralement associée à une fréquence élevée d'apparition simultanée de ces deux événements.

La présente recherche examine si les jugements phonotactiques reflètent une sensibilité particulière aux contingences phonémiques, et si celle-ci varie en fonction de la fréquence des unités composites. Bien qu'une manipulation parfaitement orthogonale de la contingence des phonèmes et de la fréquence de l'unité composite s'avère impraticable, il est possible de délimiter des unités fréquentes dont la contingence des phonèmes constitutifs est soit élevée, soit faible. De même, il existe des unités rares dont la contingence des phonèmes est soit élevée, soit faible.

2. EXPERIENCE 1 : VOYELLE-CODA

L'expérience 1 examine si la fréquence de la rime et la contingence voyelle-coda de syllabes Consonne-Voyelle-Consonne (CVC) sans signification influencent les préférences dans une tâche similaire à celle utilisée par Treiman et coll. en anglais [Tre98]. Six types de paires de syllabes étaient utilisées. Les

syllabes des quatre premiers types de paires contrastaient maximalement soit par la fréquence de la rime (F+: fréquence élevée; F-: fréquence faible), soit par la contingence de la voyelle et de la coda (C+: contingence forte; C-: contingence faible). Les contrastes de contingence étaient réalisés avec des unités fréquentes ou rares, et les contrastes de fréquence étaient réalisés avec des unités contingentes ou peu contingentes. Un cinquième type de paires contrastait des syllabes incluant une unité contingente et fréquente avec des syllabes incluant une unité peu contingente et rare. Finalement, le sixième type de paires contrastait des syllabes incluant des rimes fréquentes mais peu contingentes, avec des syllabes incluant des rimes rares mais très contingentes.

2.1 Méthode

Quatre groupes d'unités rimes ont été constitués en fonction de la fréquence et de la contingence voyelle-coda. Les rimes étaient soit contingentes et fréquentes (C+F+; ex. /aʒ/, /om/), contingentes et peu fréquentes (C+F-; ex. /uʃ/, /op/), peu contingentes et fréquentes (C-F+; ex. /yl/, /id/), ou peu contingentes et peu fréquentes (C-F-; ex. /ɑ̃l/, /ɛp/). Chaque groupe comportait six rimes. Le corpus lexical de référence pour les calculs de fréquence et les tables de contingence est la base de données psycholinguistiques Brulex [Con90] incluant approximativement 35.000 entrées lexicales. Afin d'éviter les problèmes liées aux ambiguïtés de syllabification de certains mots, les calculs n'ont portés que sur les unités rimes des syllabes finales des mots. Les calculs de fréquence et de contingence ont été réalisés de deux manières différentes, soit en fonction du nombre absolu de mots de Brulex incluant les voyelles et coda considérées (par Type), soit en fonction de la fréquence d'usage de ces mêmes mots telles que disponibles pour le langage écrit (par Token; [Imb71]). Les rimes contingentes sélectionnées se caractérisent par une fréquence de juxtaposition voyelle-coda significativement supérieure (par Type et Token) à celle prédite à partir des fréquences individuelles de la voyelle et de la coda. Les valeurs moyennes des fréquences et des contingences (coefficient Ø) pour chacun des quatre groupes apparaissent dans la Table 1. Notons que le coefficient Ø (maximum: 1) tient compte des covariations envisagées de façon bidirectionnelle; c'est-à-dire de la présence d'une coda particulière compte tenu de la présence d'une voyelle particulière, mais aussi de la présence de la voyelle compte tenu de la coda.

A partir de chacune des 24 unités rimes sélectionnées, trois syllabes CVC sans signification étaient générées par l'addition d'une consonne initiale (attaque). Les contraintes dans le choix des attaques étaient que l'unité composite CV résultante ne constitue ni une unité fréquente, ni contingente. Les 72 syllabes CVC ainsi obtenues étaient combinées de façon à constituer 36 paires de syllabes se répartissant en 6 différents types de 6 paires. Les syllabes incluses dans les quatre premiers types différaient soit par la contingence voyelle-coda (pour des unités rimes fréquentes ou rares) , soit par la fréquence de l'unité rime (pour des contingences voyelle-coda élevées ou faibles). Les syllabes en opposition dans le cinquième type de paires possédaient soit une rime fréquente et contingente, soit une rime rare et peu contingente. Finalement, les rimes des syllabes du sixième type de paire étaient soit fréquentes et peu contingentes, soit rares et contingentes. Les syllabes de chaque paire étaient appariées de façon à ce que la fréquence moyenne des trois phonèmes (telle que calculée à partir des mots monosyllabiques de Brulex) soit comparable.

Table 1: Fréquences et contingences (coefficient Ø) moyennes des unités voyelle-coda (Expérience 1) et attaque-voyelle (Expérience 2).

	Types d'unités			
	C+		C-	
Variables	F+	F-	F+	F-
Expérience 1				
Fréquence	352	24	107	3
	(124)	(12)	(38)	(2)
Contingence	0,21	0,08	0,04	0,01
	(0,17)	(0,09)	(0,02)	(0,01)
Expérience 2				
Fréquence	264	99	163	2
	(124)	(55)	(76)	(1)
Contingence	0,09	0,06	0,01	0,02
	(0,09)	(0,08)	(0,01)	(0,02)

Note: Moyennes par Type (par Token entre parenthèses). C = Contingence; F = Fréquence; + = forte; - = faible.

Les 36 paires de syllabes CVC furent enregistrées par le second auteur, et digitalisées (16 bits, 44,1 kHz). Deux ordres de présentation des 36 paires furent créés, en inversant l'ordre de présentation des paires de syllabes, ainsi que des syllabes à l'intérieur de chaque paire. Les deux listes expérimentales furent transférées sur bande DAT. Chaque paire de syllabes était présentée deux fois successivement. Chacun des 36 essais débutait par un bip sonore suivi 500 ms plus tard de la première présentation de la paire de syllabe. La seconde présentation commençait 1,5 s après la fin de la première présentation. L'intervalle de temps séparant la fin de la première syllabe du début de la seconde syllabe était de 900 ms. Pour chaque essai, les enfants testés devaient décider laquelle des deux syllabes sans signification paraissait la plus conforme aux mots de la langue française. La population examinée comprend 21 enfants de troisième année primaire de l'agglomération dijonnaise. Ils étaient testés individuellement.

2.2 Résultats et discussion

Les moyennes des préférences calculées sur les six essais (valeurs comprises entre 0 et 6) pour chaque type de paire de syllabes sont présentées dans la Table 2. Pour chaque type de paires, des tests t comparent les préférences moyennes obtenues au niveau du hasard (3). Les comparaisons significatives sont indiquées dans la Table 2.

Lorsque les syllabes à comparer varient à la fois en termes de contingence et de fréquence de la rime (C+F+ / C-F-), les préférences portent sur la syllabe possédant une rime contingente et fréquente. Cette observation est similaire à celle décrite par Treiman et coll. (Tre98) pour des syllabes anglaises. La contingence de la rime semble également guider le choix lorsque la rime est fréquente (C+F+ / C-F+) mais pas lorsqu'elle est rare (C+F- / C-F-). Cette variation de l'effet de contingence en fonction de la fréquence est tendanciel ($p = .07$). Inversement, la fréquence de la rime influence le choix lorsque les rimes sont contingentes (C+F+ / C+F-), mais pas lorsqu'elles ne sont pas contingentes (C-F+ / C-F-). La variation de l'effet de fréquence en fonction de la contingence est significative ($F(1,20) = 4.95$, $p<.05$). Finalement, aucune préférence significative n'est relevée lorsque les syllabes en opposition incluent d'une part une rime contingente mais rare, et d'autre part une rime peu contingente mais fréquente (C+F- / C-F+).

Table 2: Moyennes des préférences pour chaque type de paire de syllabes dans l'Expérience 1 (voyelle-coda) et l'Expérience 2 (attaque-voyelle).

Type de paire	Expérience 1	Expérience 2
C+F+ / C-F-	3,71 / 2,29 *	3,61 / 2,39 *
C+F+ / C-F+	3,62 / 2,38 *	2,87 / 3,13
C+F- / C-F-	2,91 / 3,09	3,70 / 2,30 *
C+F+ / C+F-	3,95 / 2,05 *	3,04 / 2,96
C-F+ / C-F-	3,00 / 3,00	4,26 / 1,74 *
C+F- / C-F+	3,24 / 2,76	3,09 / 2,91

Note: Les moyennes apparaissent dans l'ordre des syllabes tel que spécifié dans la colonne Type de paire; * différences significatives par rapport au hasard ($p<.05$). C = Contingence; F = Fréquence; + = forte; - = faible.

Il semble donc que les contingences phonémiques voyelle-coda, ainsi que les fréquences des rimes influencent toutes deux les jugements phonotactiques. Toutefois, les variations des effets de contingence en fonction de la fréquence, et des effets de fréquence en fonction de la contingence pourraient résulter des caractéristiques du matériel utilisé. Ainsi, l'observation d'un effet de contingence seulement lorsque les rimes sont fréquentes est associée à une manipulation de la contingence (coefficient Ø dans la Table 1) plus importante pour les rimes fréquentes que pour les rimes rares (différences de 0,17 vs. 0,7). Une interprétation similaire est également possible afin de rendre compte de l'observation que la fréquence de l'unité rime n'influence les préférences que lorsque la contingence voyelle-coda est élevée. Il nous est donc paru nécessaire d'examiner les effets de contingence et de fréquence avec un nouveau matériel mieux équilibré.

3. EXPERIENCE 2 : ATTAQUE-VOYELLE

L'expérience 2 consiste en une réplication exacte du plan expérimental de l'Expérience 1 excepté que les manipulations de fréquence et de contingence portent sur les unités attaque-voyelle des syllabes pour lesquelles les variations de contingence et de fréquence peuvent être contrôlées plus finement.

3.1 Méthode

Quatre groupes de six unités attaque-voyelle furent constitués en fonction des variables de fréquence et de contingence phonémique (C+F+, ex. /bu/, /di/; C-F+, ex. /my/, /pu/; C+F-, ex. /ʃɑ̃/, /nɔ/; C-F-, ex. /tœ/, /za/). Les fréquences et contingences attaque-voyelle furent déterminées à partir du corpus lexical de Brulex [Con90] en ne considérant que les syllabes initiales des mots. Les valeurs moyennes pour chacun des quatre groupes d'unités attaque-voyelle sont fournies dans la Table 1.

A partir de chacune des 24 unités attaque-voyelle, trois syllabes CVC sans signification furent générées par l'adjonction d'une consonne finale, avec la contrainte que celle-ci ne forme pas une unité contingente et/ou fréquente avec la voyelle précédente. Les autres principes de construction du matériel expérimental et la procédure utilisée sont similaires à ceux de l'Experience 1. Vingt-trois enfants de troisième année primaire participaient à l'expérience.

3.2 Résultats et discussion

Les résultats apparaissent dans la Table 2. Les analyses sont similaires à celles réalisées dans l'Experience 1. Lorsque les syllabes à comparer varient à la fois en termes de contingence et de fréquence de l'unité attaque-voyelle (C+F+ / C-F-), les préférences portent sur la syllabe possédant une unité contingente et fréquente. La contingence influence le choix lorsque l'unité est rare (C+F- / C-F-) mais pas lorsqu'elle est fréquente (C+F+ / C-F+). Cette variation significative ($F(1,22) = 4.25$, $p=.05$) de l'effet de contingence en fonction de la fréquence est inverse à celle observée dans l'Expérience 1. La fréquence de l'unité attaque-voyelle influence le choix lorsque la contingence attaque-voyelle est faible (C-F+ / C-F-), mais pas lorsqu'elle est élevée (C+F+ / C+F-). La variation de l'effet de fréquence en fonction de la contingence est significative ($F(1,22) = 9.91$, $p<.01$), et dans la direction opposée à celle observée dans l'Expérience 1. Finalement, aucune préférence significative n'est relevée lorsque les syllabes en

opposition incluent d'une part une unité attaque-voyelle contingente mais rare, et d'autre part une unité peu contingente mais fréquente (C+F- / C-F+).

4. Discussion generale

Les travaux examinant la sensibilité aux règles phonotactiques de la langue ont généralement été menés en jouant sur de larges variations fréquentielles en contrastant des suites phonémiques légales et illégales. En outre, à l'exception de l'étude de Treiman et coll. [Tre98], aucune étude n'a à notre connaissance examiné l'influence de la contingence phonémique intra-syllabique. Bien que les variables de fréquence et de contingence soient fortement corrélées, nous avons tenté de dissocier autant que possible les effets des deux variables. Dans l'ensemble, les données suggèrent que les jugements phonotactiques sont influencés par les caractéristiques fréquentielles et de contingence des unités sous-syllabiques.

Engagé dans une tâche où il doit choisir laquelle de deux syllabes « sonne » le plus comme les mots de sa langue, le sujet pourrait aussi être guidé par l'existence de mots ressemblant aux syllabes sans signification (voisins phonologiques). Une opérationnalisation de la notion de voisinage est de considérer comme voisin phonologique d'une syllabe tout mot de même nombre de phonèmes différant d'un phonème au maximum. Il est probable que le nombre (N) de voisins phonologiques varie en fonction de la fréquence et de la contingence des unités manipulées, surtout lorsque les syllabes incluent des rimes fréquentes. En effet, les mots voisins ont généralement la rime en commun [Pee97]. Des analyses post-hoc ont examiné si les préférences exprimées étaient fonction du N. Les résultats indiquent que la différence de N entre les syllabes de chaque paire corrèle avec la différence de préférences pour l'Expérience 1 ($r = 0{,}39$; $p < .05$), mais pas pour l'expérience 2. Ainsi, bien qu'une contamination des résultats de l'expérience 1 ne puisse être exclue, la similarité globale des syllabes sans signification avec des mots réels ne semble pas pouvoir expliquer les résultats de l'Expérience 2.

Finalement, les différences de résultats observées pour les unités voyelle-coda et attaque-voyelle pourraient refléter de subtiles variations dans la sensibilité aux contingences phonémiques en fonction de la position interne à la syllabe. L'observation que la contingence attaque-voyelle n'influence pas les choix lorsque les unités sont fréquentes peut résulter d'une sensibilité progressivement asymptotique aux contingences phonémiques en fonction de la fréquence d'exposition de l'unité attaque-voyelle. Plus surprenant est le patron inverse des résultats recueillis pour l'unité voyelle-coda. Plusieurs auteurs suggèrent que la rime constituerait une unité plus cohésive que l'unité attaque-voyelle. Cette cohésion résulterait d'une sonorité plus proche entre la coda et la voyelle qu'entre l'attaque et la voyelle [Tre84]. Si cette plus grande cohésion de la rime rend l'individualisation des phonèmes plus ardue, la sensibilité aux contingences phonémiques pourrait nécessiter une fréquence d'exposition plus importante.

Bibliographie

[Sei97] Seidenberg, M. S. (1997, « Language acquisition and use: Learning and applying probabilistic constraints », Science, Vol. 275, pp. 1599-1603.

[Sha95] Shanks, D. R. (1995). The psychology of associative learning. Cambridge: Cambridge University Press.

[Fra96] Frauenfelder, U. H., Content, A., & Peereman, R. (1996), « L'utilisation de bases de données lexicales en psycholinguistique: Applications à la reconnaissance des mots », Lexique et Communication Parlée, GDR-PRC, Toulouse.

[Saf96] Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996), « Statistical learning by 8-Month-Old Infants », Science, Vol. 274, pp. 1926-1928.

[Jus97] Jusczyk, P. W. (1997). The discovery of spoken language. MIT Press.

[Per81] Perruchet, P. (1981). « Les conditions d'établisement des liaisons conditionnelles: Quelques perspectives récentes », Journal de Psychologie, Vol. 2-3, pp. 298-319.

[Tre98] Treiman, R., Kessler, B., Knewasser, S. Tincoff, R., & Bowman, M. (1998), « Phonotactic patterns in English words and speakers'sensitivity to these patterns », Papers in Laboratory Phonology V, Cambridge University Press.

[Con90] Content, A., Mousty, P., & Radeau, M. (1990). « Brulex. Une base de données lexicales informatisée pour le français écrit et parlé », Année Psychologique, Vol. 90, pp. 551-566.

[Imb71] Imbs, P. (1971), « Etudes statistiques sur le vocabulaire français. Dictionnaire des fréquences », Paris, Marcel Didier.

[Pee97] Peereman, R., & Content, A. (1997). « Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space », Journal of Memory and Language, Vol. 37, pp. 382-410.

[Tre84] Treiman, R. (1984). « On the status of final consonant clusters in English syllables », Journal of Verbal Learning and Verbal Behavior, Vol. 23, pp. 343-356.

FACILITATION INITIALE EN AMORÇAGE PHONOLOGIQUE, EFFET SYLLABIQUE OU PHONÉMIQUE ?

Elsa Spinelli & Juan Segui

Laboratoire de Psychologie Expérimentale
28 rue Serpente, 75006 Paris
Tél.: 40 51 98 65 - Fax: 40 51
e-mail: spinelli@idf.ext.jussieu.fr

ABSTRACT

Three experiments using cross modal phonological priming were carried out to determine the functionnal units involved in lexical access. Results of experiment 1 revealed a facilitation for the beginning overlap condition. Experiment 2 (lexical decision task) and 3 (naming) focused on initial overlap only and the degree of phonological overlap and the syllabic segmentation were varied as separate factors. The results showed that CVC primes gave rise to shorter reaction times than primes that corresponded to the syllabic structure of the target.

1. INTRODUCTION

Le processus global de compréhension du langage passe par la reconnaissance des mots. Pour ce faire, l'auditeur doit identifier des formes de mots dans l'information acoustique perçue afin d'accéder aux représentations lexicales en mémoire. C'est précisément l'organisation du lexique interne ainsi que les processus impliqués dans l'accès aux représentation de mots que nous étudions.

A la base de nos recherches se trouve l'idée que l'information lexicale en mémoire est organisée sur la base de relation phonologique.

Le paradigme d'amorçage est une technique utilisée pour explorer l'organisation du lexique mental ainsi que pour déterminer les unités intervenant dans l'accès lexical. L'amorçage phonologique consiste à présenter une amorce qui partage un certain nombre de phonèmes avec la cible et voir si le recouvrement phonologique facilite ou inhibe le traitement ultérieur de la cible.

Les différents modèles de reconnaissance des mots parlés font des prédictions différentes quant aux effets du recouvrement phonologique. Selon le modèle de la Cohorte de Marslen-Wilson [Mar78], un auditeur sélectionne à l'aide des deux ou trois phonèmes qu'il perçoit, un premier ensemble de candidats. Chaque phonème entrant élimine de la cohorte les candidats qui ne correspondent plus au signal et un mot est reconnu lorsqu'il ne reste qu'un seul candidat. Cependant les membres de la cohorte restent activés un certain laps de temps avant de décliner en activation. Ce modèle prédirait donc un effet facilitateur du traitement des cibles, si la cible fait partie de la cohorte générée par l'amorce, du fait de l'activation résiduelle des membres de la cohorte.

Les études précédentes portant sur la question du contexte phonologique se sont intéressées à 2 types de recouvrement : le recouvrement initial (Blague-Blame) et le recouvrement final (Trève-Grève). En ce qui concerne le recouvrement final, on observe le plus souvent un effet de facilitation mais uniquement lorsque amorces et cibles sont présentées dans la même modalité, en l'occurrence auditive [Rad 95]. Cet effet de facilitation ne persiste pas en situation d'intermodalité.

En ce qui concerne le recouvrement initial, deux types d'effet ont été observés en fonction du degré de recouvrement phonologique. Lorsque amorces et cibles ne partagent que le phonème initial, on observe un effet de facilitation du traitement des cibles [Slo 92], mais lorsque amorces et cibles partagent les deux ou les trois phonèmes initiaux, c'est un effet d'inhibition qui est observé [Ham 96]. Contrairement à l'effet de facilitation observé généralement pour le recouvrement final, cet effet d'inhibition du recouvrement initial est observé non seulement en intramodalité mais également en intermodalité. La logique de la séparation entre amorçage intramodalité et intermodalité est qu'on suppose généralement que les processus prélexicaux sont spécifiques à la modalité d'entrée alors que la situation d'intermodalité ne peut que refléter des processus opérant au niveau du lexique. Si en effet la présentation auditive d'une amorce influence le traitement d'une cible présentée visuellement, cet effet ne peut se situer qu'au niveau lexical. Ainsi, les auteurs interprètent ces résultats en terme de mécanismes de compétition entre voisins phonologiques.

Jusqu'à présent, toutes les études ont été menées sur des amorces et des cibles monosyllabiques. Dans cette présente recherche, nous avons pensé que travailler sur des mots bisyllabiques présentait un intérêt supplémentaire. Premièrement, cela nous permet de déterminer quelle est la nature des unités fonctionnelles qui permet le contact avec les entrées lexicales. Enfin, cela nous permet de contraster des effets phonémiques des effets syllabiques.

2. EXPERIENCE 1 : EFFET DE FACILITATION DU RECOUVREMENT INITIAL :

2.1Methode :

Sujets : 30 sujets de langue maternelle française ont participé à l'expérience dans le cadre de leur Licence de Psychologie. Tous avaient une vision normale ou corrigée et n'ont reporté aucun trouble particulier de l'audition.

Stimuli :

Mots : 36 mots bisyllabiques ont été sélectionnés dans la base de donnée TLF. Chacun de ces mots contenait deux mots enchâssés (ex : Verveine). A ces 36 mots cibles, étaient associés 108 mots amorces, 36 d'entre eux entretenaient une relation de recouvrement initial avec la cible (i.e. correspondaient à la première syllabe du mot Cible (ex Ver)), 36 d'entre eux entretenaient une relation de recouvrement final avec la cible (i.e. correspondaient à la dernière syllabe du mot Cible (ex Veine)) et 36 d'entre eux n'entretenaient aucune relation avec la cible (ex Cap).

Non Mots : 36 non-mots bisyllabiques ont été créés de façon à contenir deux mots enchâssés (ex : Murdose). A ces 36 non-mots cibles, étaient associés 108 mots amorces monosyllabiques entretenant les mêmes relations (début, fin et contrôle) que dans le cas des mots. Les non mots servent donc d'essais attrapes afin que le sujet ne puisse pas faire un jugement de lexicalité sur la seule base d'un lien amorce-cible.

Fillers : 120 items de remplissage bisyllabiques ont été insérés dans les listes (60 mots et 60 non-mots). A ces 120 cibles de remplissage ont été associés 120 mots amorces monosyllabiques qui n'entretenaient aucune relation avec les cibles de remplissage.

Procédure : Enregistrement des stimuli sonores :

Les items sont enregistrés sur un DAT (magnétophone digital) dans une pièce insonorisée, par un locuteur féminin de langue maternelle française. Le matériel est échantillonné à 44 kHz puis transféré sur un micro-ordinateur à une fréquence de 22,05 kHz, en 16 bits et en monophonie grâce au logiciel " Wave Studio " pour Sound Blaster.

Passation de l'expérience :

La passation s'effectue individuellement. Le sujet commence par entendre une amorce auditive présentée dans un casque. A la fin du mot amorce, et après un délai de 50 ms (Isi), un item cible apparaît visuellement au centre de l'écran d'un ordinateur. Cet item reste affiché à l'écran jusqu'à la réponse du sujet.

La tâche du sujet est une tâche de décision lexicale : il doit décider le plus rapidement possible si la séquence de lettres présentées à l'écran constitue un mot de la langue française ou pas. Il donne sa réponse en appuyant sur l'un des deux boutons mis à sa disposition (mot / non-mot). L'horloge de l'ordinateur est déclenchée à l'apparition de l'item cible et s'arrête à la réponse du sujet. Afin que chaque sujet soit confronté à toutes les conditions expérimentales mais ne voie pas deux fois le même item cible, la répartition des mots cibles a été contrebalancée dans trois listes. Lors de la passation, les items sont présentés au sujet dans un ordre aléatoire.

2.2 Résultats :

Les analyses de variance effectuées sur ces données révèlent un effet global d'amorçage ($F1(2,58)=9.67$; $p<.001$, $F2(2,70)=8,18$ $p<.001$), ainsi qu'un effet de facilitation du partage initial ($F1(1,29)=14.16$; $p<.001$; $F2(1,35)=5.36$; $p<.025$).

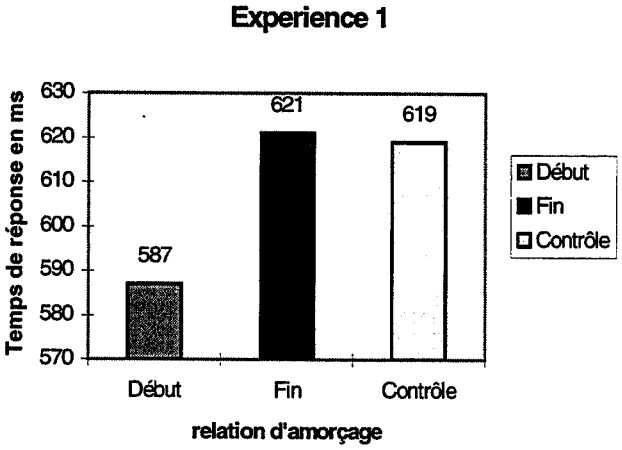


Figure1: Graphique des temps de réaction (en millisecondes) pour les cibles dans les 3 conditions d'amorçage.

2.3Discussion:

Cette première expérience en amorçage auditif-visuel a montré que la présentation auditive d'une amorce comme « ver » facilite le traitement ultérieur d'une cible présentée visuellement comme « verveine » par rapport à une amorce non reliée phonologiquement comme « bac». Ainsi, on observe un effet de facilitation du recouvrement phonologique initial entre amorces et cibles, alors que le recouvrement phonologique final (comme dans le cas de « veine »-« verveine ») ne donne pas lieu à une facilitation du traitement des cibles.

Ces résultats vont dans le sens du modèle de reconnaissance des mots parlés « Cohorte » [Mar 78]. Dans le cas du recouvrement initial, la présentation de l'amorce ouvre une cohorte de candidats qui contient entre autres, la cible (par exemple le mot « veveine » fait partie de la cohorte générées par le mot « ver »). Du fait de l'activation résiduelle des membres de la cohorte générée par l'amorce, le traitement de la cible

se trouve facilité. En ce qui concerne le recouvrement final, nous observons une absence de facilitation. Ce résultat est en accord avec la littérature qui ne reporte des effets de recouvrement final (ou rime) que lors d'une situation d'intramodalité. Ces effets sont supposés avoir lieu au niveau prélexical et ne résistent pas à la situation d'intermodalité.

3. EXPERIENCE 2 : EFFET SYLLABIQUE OU EFFET PHONEMIQUE ?

Dans cette expérience, on s'intéresse au recouvrement initial uniquement et au découpage syllabique. L'expérience 1 nous a permis de mettre en évidence un effet de facilitation du recouvrement phonologique initial. Nous avons constaté que le partage des premiers phonèmes donne lieu à un effet de facilitation, cependant ce recouvrement phonologique correspondait également au découpage syllabique de la cible (« ver » est la première syllabe de « verveine »). Donc nous avons voulu savoir si l'effet facilitateur du recouvrement initial obtenu précédemment est seulement dû au partage de phonèmes entre amorces et cibles, ou bien si le fait que le partage initial corresponde également au découpage syllabique entre en jeu. Ainsi, cette expérience avait pour but d'inférer quelles sont les unités fonctionnelles qui servent à contacter le lexique et plus précisément, de voir si certaines composantes comme la syllabe sont impliquées dans l'établissement du contact entre l'input et l'entrée lexicale correspondante. Pour cela, nous avons fait varier le degré de recouvrement phonologique initial en fonction du découpage syllabique des cibles.

3.1Methode :

Sujets : 20 sujets de langue maternelle française ont participé à l'expérience dans le cadre de leur Licence de Psychologie. Tous avaient une vision normale ou corrigée et n'ont reporté aucun trouble particulier de l'audition.

Stimuli :

Mots : 14 paires de mots de syllabe initiale CVC ou CV comme Virgule -Virage appariées en fréquence ont été sélectionnées dans la base de donnée TLF.

Non Mots : 14 paires de syllabe initiale CVC ou CV comme Barbin -Barine ont été crées.

FIllers : 360 items de remplissage bisyllabiques ont été insérés dans les listes (180 mots et 180 non-mots). A ces 360 cibles de remplissage ont été associés 360 mots amorces monosyllabiques qui n'entretenaient aucune relation avec les cibles de remplissage.

Procédure : La procédure d'enregistrement des stimuli, la passation de l'expérience et la tâche du sujet sont les mêmes que dans l'expérience 1.

Donc chaque sujet est confronté à quatre situations expérimentales, 1) une condition amorce CV/cible CV (ex : BA-BALANCE), une condition CVC / CVC (ex : BAL-BALCON), une condition CVC / CV (ex : BAL-BALANCE), une condition CV / CVC (ex : BA-BALCON). Dans les deux premiers cas, l'amorce est en accord avec le découpage syllabique de la cible alors que dans les deux dernières conditions, l'amorce ne correspond pas au découpage syllabique de la cible.

Si l'effet de facilitation du recouvrement initial obtenu précédemment était dû à un accord syllabique entre amorces et cibles, alors on devrait s'attendre à un effet d'interaction croisée, c'est à dire une plus grande facilitation dans le cas de « ba » « balance » et « bal » « balcon » que dans le cas de « bal » « balance » et « ba » « balcon ». Par contre, si seul le degré de recouvrement est un facteur pertinent, on devrait s'attendre à de plus grands effets de facilitation lorsque le recouvrement est maximal, quel que soit le découpage syllabique des cibles (donc à une plus grande facilitation des amorces « bal » par rapport aux amorces « ba »).

3.2 Résultats :

Les analyses de variance effectuées sur ces données révèlent un effet global d'amorçage, les amorces CVC donnant lieu à un amorçage plus important que les amorces CV ($F1(1\text{-}18)=14.75$; $p<001$; $F2(1\text{-}26)=7$; $p<.01$) ainsi qu'un effet global de la structure syllabique des cibles, les cibles CVC donnant lieu à des temps de réponse plus courts ($F1(1\text{-}18)=8.45$; $p<.005$; $F2(1\text{-}26)<1$). L'interaction entre ces eux facteurs n'est pas significative ($F1(1\text{-}18)<1$)

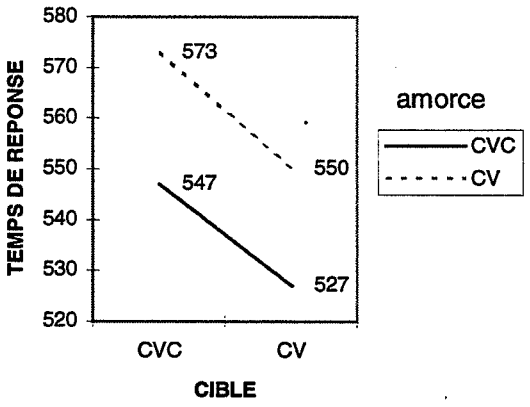


Figure 2: Graphique des temps de réaction (en millisecondes) pour les cibles CVC et CV dans les 2 conditions d'amorçage CVC et CV.

3.3Discussion :

Les résultats sont en faveur de la deuxième interprétation, les amorces CVC (comme « bal ») induisent une facilitation plus importante que les

amorces CV (comme « ba »). Les unités syllabiques ne rentreraient donc pas en compte pour déterminer la cohorte de candidats activée par une séquence de phonèmes. Ainsi, une amorce comme « vol » active une cohorte de candidats qui comprend notamment des mots comme « voltige », « volcan » mais également « voleur » ou « volonté ». Cependant, afin de confirmer cette hypothèse, nous avons voulu vérifier les résultats avec une tâche de production, supposée plus sensible aux effets syllabiques.

4. EXPERIENCE 3 :

4.1Methode :

Sujets : 20 sujets de langue maternelle française ont participé à l'expérience dans le cadre de leur Licence de Psychologie. Tous avaient une vision normale ou corrigée et n'ont reporté aucun trouble particulier de l'audition.

Stimuli : Les stimuli étaient les mêmes que ceux de l'expérience précédente. Seuls les mots ont été gardés.

Procédure : La procédure est la même que dans l'expérience précédente mise à part la tâche du sujet qui est une tâche de lecture à haute voix, le sujet devant lire le plus rapidement possible les items présentés à l'écran sur l'ordinateur. Les latences de lecture (intervalle qui sépare la présentation de la cible et le début de la réponse du sujet) sont mesurées à l'aide d'une clef vocale.

4.2 Résultats :

Comme pour l'expérience précédente, les analyses de variance effectuées sur ces données révèlent un effet global d'amorçage, les amorces CVC donnant lieu à un amorçage plus important que les amorces CV ($F1(1-18)=12.55$; $p<.002$). De même, l'interaction entre le facteur structure des amorces et le facteur structure des cibles n'est pas significative ($F1(1-18)<1$; ns).

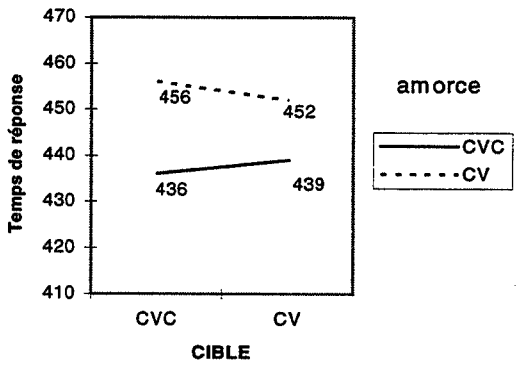


Figure 3: Temps de réaction (en ms) pour les cibles CVC et CV dans les 2 conditions d'amorçage CVC et CV.

4.3Discussion :

Les résultats répliquent ceux trouvés avec une tâche de décision lexicale. Il semble alors que la syllabe n'est pas une unité fonctionnelle dans le processus d'accès au lexique. Cependant, des expériences précédentes ont montrés qu'une incompatibilité orthographique entre amorce et cible conduisait à la disparition des effets d'amorçage en situation d'intermodalité mais pas en situation intramodale [Spi 97]. Ces résultats nous conduisent à repenser notre interprétation sur les effets obtenus sur le découpage syllabique dans la mesure où la plupart des items CVC (donc ceux qui induisaient une facilitation plus importante) étaient pour la plupart d'entre eux, compatibles orthographiquement, (par exemple: « bal » « par » « car » dont la réalisation orthographique est contenue telle qu'elle dans la cible). Par contre, les items CV étaient pour la plupart non compatibles orthographiquement. Les résultats obtenus sur la syllabe mériteraient donc d'être examinés. Afin de préciser la nature des effets obtenus dans nos expériences nous sommes en train de mener cette expérience en intramodalité (où amorces et cibles sont présentées auditivement). Si la nature de nos effets obtenus sur la syllabe est à repenser en terme de compatibilité orthographique, le fait de court-circuiter l'effet de l'orthographe en présentant amorces et cibles en modalité auditive [Spi 97] nous renseignera sur le caractère phonémique ou syllabique des effets de facilitation dû recouvrement initial obtenus en tout premier lieu.

BIBLIOGRAPHIE

[Ham 96] Hamburger, M. et Slowiaczeck, L.M. (1996). "Phonological priming reflects lexical competition". Psychonomic bulletin & review, 3 (4), 520-525.

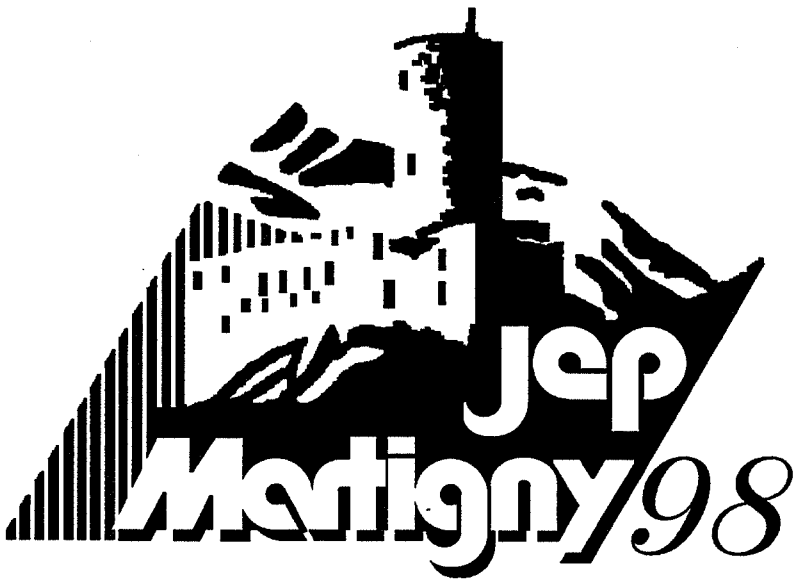
[Mar 78] Marslen-Wilson, W. D., & Welsh, A.(1978). "Processing interaction and lexical access during word recognition in continuous speech". Cognitive Psychology, 10, 29-63..

[Rad 95] Radeau, M.; Morais, J. et Segui, J. (1995). "Phonological priming between monosyllabic spoken words". Journal of experimental psychology : Human perception and performance, 21(6), 1297-1311.

[Slo 92] Slowiaczeck, L.M. et Hamburger, M. (1992). "Prelexical facilitation and lexical interference in auditory word recognition". Journal of experimental psychology: Learning, Memory and Cognition, 18 (6), 1239-1250.

[Spi 97] Spinelli, E. (1997). "Amorçage phonologique et reconnaissance des mots parlés". Poster présenté aux IIèmes RJC en parole. La Rochelle, 14-15 Nov 1997.

Conférence invitée



Dynamique des systèmes phonologiques

Aspects formels et expérimentaux

Didier DEMOLIN

Laboratoire de phonologie, Université libre de Bruxelles
LPL-Aix-en-Provence et DDL-Lyon2
email : ddemoli@ulb.ac.be

Abstract

This paper examines the dynamics of phonological systems from different point of views. First the main dynamic properties of articulatory phonology are evoked. Then several properties of self-organisation in phonology are considered, i. e. the role of self-organisation in the emergence and in the structuration of phonological systems. Diachronic cases concerning vowels and clicks are then examined to show how important these factors are to understand the evolution of phonological systems. The experimental part of the paper is devoted to study aspects of the control and regulation of phonological systems primarily based on the interaction of articulatory and aerodynamic factors. Finally a proposition to consider a formalism borrowed frrom biology is presented.

1. Introduction

Une partie importante de la recherche en phonologie de ces quarante dernières années à été consacrée à la formalisation des représentations des unités et des processus qui déterminent les systèmes phonologiques, ainsi qu'à la recherche des primitifs qui caractérisent ces sytèmes. De nombreux modèles ont vu jour dans les années qui ont suivi la publication du 'Sound Pattern of English' de Chomsky et Halle en 1968 [Cho68] (pour des détails voir par exemple [Dur90], [Gol95], [Cla94]). Les travaux qui ont été menés dans cet esprit ont consacré l'essentiel de leurs efforts à trouver les représentations adéquates pour décrire les éléments des systèmes phonologiques et à trouver les contraintes qui gèrent leur structure. L'introduction du geste articulatoire et du modèle de la phonologie articulatoire par Browman et Goldstein, [Bro89], [Bro92] a marqué un pas important vers l'étude de la dynamique des systèmes phonologiques. En effet, ce modèle dérivé de la dynamique des tâches [Sal89] ne considère plus le segment ou le trait comme un primitif en phonologie, mais le geste articulatoire, qui est à la fois une unité abstraite de contraste et une unité d'action douée d'un aspect spatio-temporel. Les principales caractéristiques du modèle de Browman et Goldstein seront examinées. Ensuite d'autres aspects de la dynamique des sytèmes phonologiques seront abordés. D'une part, l'étude des méthodes déductives qui permettent de rendre compte de la génération de certaines structures, d'autre part le rôle que joue l'auto-organisation dans la genèse des sytèmes et dans leur fonctionnement collectif, et pour terminer, l'importance des facteurs diachroniques pour comprendre la dynamique des systèmes. Une des conséquences importantes d'une approche dynamique de la phonologie concerne le problème du contrôle et de la régulation du système et de ses éléments. Cette question sera abordée par l'examen des moyens d'investigation qui sont à la disposition des phonologues lorsqu'ils essayent de comprendre l'interaction et les rétroactions qui existent entre les différents composants de la production de la parole. Quelques cas concrets d'interaction et de rétroaction entre les niveaux aérodynamiques, articulatoires et acoustiques seront ensuite abordés. La dernière partie de l'exposé sera consacrée au problème du choix des variables et de celui d'un système formel qui permet de décrire la dynamique des systèmes phonologiques. Cet exposé cherche à présenter, sans chercher à l'exhaustivité, quelques aspects de la dynamique des systèmes phonologiques; la relation de la phonologie avec les différents composants de la production de la parole; les liens entre la phonologie et d'autres approches de l'étude de la parole ainsi que quelques propositions pour la formalisation de la dynamique des systèmes phonologiques.

2. Phonologie

La phonologie peut se définir comme la description et l'explication des systèmes sonores des langues naturelles, ou encore comme l'étude du comportement et de la structure logique et fonctionnelle des sons de la parole. Si la phonologie est comprise de cette manière, il est alors raisonable de penser que son but peut être atteint en se fondant sur des modèles qui se basent : sur des principes qui relient la forme du conduit vocal et la sortie acoustique, quelques principes aérodynamiques connus et les principes sur lesquels se fondent notre système auditif pour extraire l'information du signal acoustique. Une telle conception de la phonologie ne considère pas la discipline comme autonome par rapport à la phonétique, mais considère plutôt que ces deux disciplines traitent du même domaine (l'univers de la parole) et doivent être intégrées. Un argument important en faveur de cette approche a été avancé par Ohala, [Oha90] : *'for the sake of explaining*

natural sound patterns there are advantages to representations using phonetic primitives. One advantage is that a few well-chosen primitives go a long way' (voir aussi [Oha86] et [Oha95] pour plus de détails sur la question). Rousselot, le père de la phonétique expérimentale, considérait déjà que les termes phonétique et phonologie désignaient le même objet. Le terme phonétique fut choisi dans le titre de ses principes, parce que phonologie, *'..., avec notre transcription peut signifier "la science du meurtre" (φονος)*' [Rou08].

Dans l'univers de la parole, la phonologie s'occupe du comportement de la parole, des représentations des mots et de la connaissance que les locuteurs ont des relations entre les mots. La phonétique étudie comment les sons de la parole sont produits et perçus. La phonétique est supposée être la partie de l'univers de la parole qui implémente les signaux de contrôle du composant phonologique. De ce point de vue, la phonologie peut être considérée comme une discipline expérimentale, puisque les hypothèses qu'elle propose peuvent faire l'objet d'observations instrumentales et d'expérimentations de différents types. Cette conception de la phonologie n'est cependant pas partagée par de nombreux linguistes pour qui la phonologie se distingue clairement de la phonétique. En effet, certains phonologues considèrent, que la tâche de la phonologie est de déterminer la structure et la représentation des systèmes qui se rencontrent dans les langues, et que l'aspect expérimental n'a que peu à voir avec leur travail. Il n'est pas inutile de rappeler à cet effet les remarques que faisait l'Abbé Rousselot au début du siècle, dans ses 'Principes de phonétique expérimentale'. *"...les procédés des sciences expérimentales sont assez étrangers aux linguistes. Une sorte de terreur superstitieuse s'empare d'eux dès qu'il s'agit de toucher au mécanisme le plus simple. Il fallait donc leur montrer que la difficulté est moindre qu'ils ne se la figurent et leur faire entrevoir le champ immense que l'expérimentation ouvre devant eux* " [Rou08]. Fort heureusement, le rejet de l'expérimentation comme source de vérification des hypothèses de travail semble perdre du terrain, comme en témoignent notamment les travaux issus du courant de la phonologie de laboratoire [Kin90], [Doc92], [Kea94], [Con95] et les travaux menés dans la lignée de Ohala et Jaeger [Oha86] et [Oha95] qui proposent de traiter la phonologie comme une discipline expérimentale.

Selon Popper, ce qui caractérise une discipline scientifique, ce ne sont pas ses méthodes ou ses théories, mais plutôt les questions qu'elle pose [Pop59]. Quelques unes des questions de base de la phonologie sont les suivantes : Comment la parole est-elle apparue? Pourquoi y a-t-il une telle diversité dans la parole? C'est-à-dire entre les communautés linguistiques et entre les générations d'une même communauté linguistique. Quelle est la structure physique de la parole? Comment est-elle produite? Comment est-elle perçue? Comment la parole est-elle représentée dans le cerveau? Comment une personne apprend-elle à prononcer sa première langue, une deuxième? Comment pouvons-nous communiquer efficacement dans des conditions hostiles? (dans un environement bruité, sur de grandes distances) Comment arrivons-nous à mémoriser la parole? Comment pouvons-nous améliorer les déficits de la parole? Comment le sens en vient-il à être associé avec les sons du langage? Comment une langue déterminée distingue-t-elle les sons? Pourquoi y a-t-il d'évidentes régularités dans les suites de sons à l'intérieur des syllabes? L'ensemble de ces questions déterminent l'objet que la phonologie partage avec d'autres disciplines; la psychologie cognitive, les neurosciences, la physique et les technologies de la parole.

Avant d'aborder la question de la dynamique des sytèmes phonologiques, un dernier point important mérite encore d'être soulevé; celui de la relation des composants du système de production de la parole avec la phonologie. On peut ainsi légitiment se poser la question de savoir quelle relation il peut y avoir entre le système phonologique et les composants aérodynamiques, neuromusculaires et acoustiques de la production de la parole. Cette relation se manifeste essentiellement au travers des différents aspects des processus de rétroaction qui existent en production de la parole. On peut par exemple observer que des modifications des conditions aérodynamiques dans le conduit vocal ont des effets observables sur certains gestes articulatoires, et que des changements acoustiques ont des effets sur les articulateurs et même sur certaines commandes motrices. Ces phénomènes de rétroaction sont fondamentaux pour comprendre la stabilité et la dynamique des systèmes phonologiques.

3. Dynamique

Avant de commencer à discuter de quelques uns des problèmes soulevés plus haut, il est utile d'introduire la notion de système dynamique que nous adopterons ici. La définition de système dynamique que nous adopterons est, simplement, celle d'un système dont le comportement à un moment donné dépend de son état à un moment antérieur. Un système dynamique est donc une équation ou un ensemble d'équations stipulant l'évolution dans le temps d'une variable x. Dans le cas de la parole, il s'agit de l'évolution d'une variable définie comme un geste articulatoire. Ce que l'on cherche à savoir, c'est comment les gestes changent d'un instant à l'autre lorsque les paramètres de contrôle varient.

3.1. Phonologie articulatoire

Comme nous l'avons mentionné plus haut, la phonologie articulatoire est dérivée du modèle de la dynamique des tâches dont le but principal est l'étude des mouvements, de leurs compensations et du temps des compensations [Sal86], [Sal89]. Avant d'aborder les principaux concepts de la phonologie articulatoire, quelques principes de base de la dynamique des tâches méritent d'être rappellés (pour une introduction plus détaillée à la dynamique des tâches voir [Haw92]). Les premiers principes qu'il convient

d'appréhender clairement sont les notions de tâche et d'objet. Une tâche est un geste qui implique le contrôle d'un objet, un objet est la variable de la tâche, la variable du conduit en parole. Une tâche se définit dans l'espace. Dans la parole, les mouvements sont caractérisés par des gestes discrets et abstraits. Un geste est défini comme une famille de mouvements qui sont fonctionnellement équivalents pour atteindre un même but. Il convient également de distinguer le geste du mouvement; un geste est défini par rapport à un but ou à une tâche, un mouvement n'est pas lié à une tâche spécifique. Une ou plusieurs variables contribuent à un même geste. Pour la parole, les gestes sont définis en termes de constriction dans le conduit vocal; le comportement d'une variable du conduit est vu par la dynamique des tâches comme un mouvement vers une cible spatiale. Ce type de tâche est appellée tâche du point attracteur. Pour chaque variable du conduit, la tâche est modélisée par une équation différentielle qui décrit le comportement d'une masse connectée à un ressort et à un amortisseur. Les mouvements de la masse représentent les changements de valeur de la variable du conduit. Le ressort tire la variable du conduit vers la valeur de sa cible. La position de repos ou d'équilibre du ressort représente la cible pour la variable du conduit et la masse bouge parce que le ressort essaye d'atteindre la position d'équilibre. Le mouvement du ressort est un mouvement amorti, l'amortissement est critique pour les gestes non-laryngaux (les gestes laryngés sont non-amortis). Cela veut dire que la masse n'oscille pas sinusoïdalement mais suit une trajectoire asymptotique vers la position d'équilibre. Pour générer les trajectoires des mouvements, il faut donc connaître, pour chaque variable pertinente, l'état instantané du système, la position de la nouvelle cible et les valeurs de raideur du ressort et d'amortissement du système. Le type d'équation différentielle qui est utilisée pour modéliser le mouvement dépend des caractéristiques de ce mouvement, pour les gestes de la parole il se définit par l'équation suivante :

$$m\ddot{x} + b\dot{x} + k(x - x_0) = 0$$

avec
- m = masse associée à la variable de la tâche
- b = amortissement du système
- k = raideur du ressort
- x_0 = position d'équilibre du ressort (la cible)
- x = valeur instantanée de la variable de la tâche (position du ressort)
- $\dot{x}$ = vélocité instantanée de la variable du conduit (tâche)
- $\ddot{x}$ = accélération instantanée de la variable de la tâche (tâche)
- $x - x_0$ = déplacement instantané de la variable de la tâche

Par convention la masse m et le terme d'amortissement b sont définis comme des constantes:

$$m = 1.0 \quad \text{et} \quad \frac{b}{2[mk]^{1/2}} = 1.0$$

Par conséquent, seulement k et x_0 doivent être spécifiés.

Le modèle de phonologie articulatoire tel qu'il a été proposé par [Bro86], [Bro89], [Bro90] est le premier et actuellement le seul modèle qui introduit explicitement un aspect dynamique en phonologie. Pour Browman et Goldstein, les gestes sont les unités de base de la phonologie, les atomes de la description. Les gestes sont des caractérisations abstraites d'événements articulatoires chacun ayant une durée intrinsèque; ce sont les unités de base de contraste dans les items lexicaux aussi bien que des unités d'action. Dès lors, la phonologie peut être vue comme un ensemble de relations parmi des événements physiques réels, une caractérisation des systèmes et des modèles dans lesquels ces événements, les gestes, entrent [Bro90]. Notons encore que les gestes ne correspondent pas à des traits ou à des segments. La différence essentielle entre traits et gestes est que le trait est un descripteur et non un primitif. Les traits sont des propriétés des segments qu'ils définissent [Fow80]. Dans le modèle de Browman et Goldstein un segment est défini comme une constellation de gestes. Comme caractérisation d'événements physiques, les gestes se déroulent dans le temps et dans l'espace; les gestes et l'organisation gestuelle peuvent être utilisés pour capturer à la fois l'information catégorielle et gradiente. La figure 1 montre les variables du modèle phonologie articulatoire telles qu'elle sont proposées dans le modèle de Browman et Goldstein.

LP : Lip protrusion : protrusion des lèvres	Lèvres inférieures et supérieures, mâchoires	LPS
LA : Lip Aperture : aperture des lèvres	Lèvres inférieures et supérieures, mâchoires	
TTCL : Tongue Tip Constriction Location : Lieu de constriction de l'apex	Apex, dos de la langue, mâchoires	TT
TTCD : Tongue Tip Constriction degree : Degré de constriction de l'apex	Apex, dos de la langue, mâchoires	
TBCL : tongue Body Constriction Location : Lieu de constriction du dos de la langue	Dos de la langue, mâchoires	TB
TBCD : Tongue Body Constriction Degree Degré de constriction de l'apex	Dos de la langue, mâchoires	
VEL : VELic aperture Ouverture du voile du palais	Vélum	VEL
GLO : GLOttal aperture Aperture glottale	Glotte	GLO

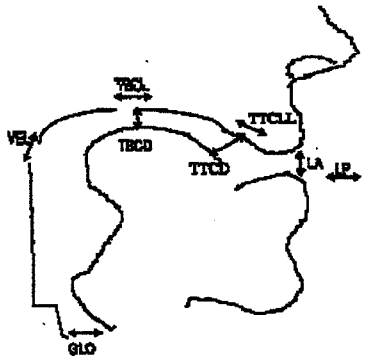


Figure 1: Définition des variables du tractus et articulateurs mis en oeuvre dans le modèle de Browman et Goldstein [Bro89].

Dans ce modèle, les gestes sont représentés sous la forme de partitions gestuelles telles que celle qui illustre le mot natte ([nat]) en français, à la figure 2. Le modèle de Browman et Goldstein permet de montrer un lien direct et naturel entre des notions qui étaient considérées soit sous un angle strictement phonologique (le contraste entre items lexicaux) soit sous un angle strictement phonétique (la modification temporelle d'un geste articulatoire). Par exemple, les pointillés qui prolongent le geste du voile à la figure 2 indiquent une variation dans la fermeture du voile.

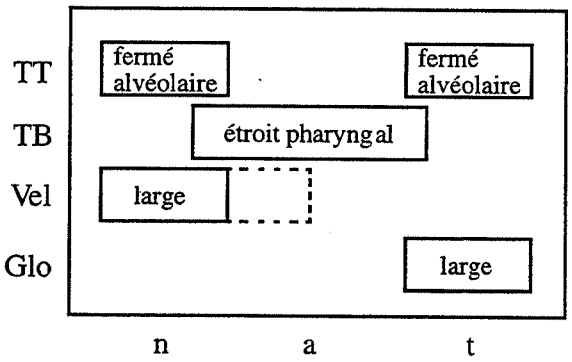


Figure 2: Partition gestuelle du mot 'natte' [nat]. Les boites correspondent aux gestes individuels pour un même niveau articulatoire. Sur chaque boite apparaît la valeur du geste pour le degré et le lieu de constriction. L'extension horizontale d'une boite donne le temps d'activation du geste.

Cette fermeture tardive rend compte de la nasalisation de la voyelle qui suit. Un geste de fermeture anticipatoire du voile permet de rendre compte de l'apparition de consonnes prénasalisées [mb] à partir de nasales simples ou géminées ([m > mb]), comme cela est observé dans les dialectes de l'Italie du sud [Rem84]. L'introduction du geste comme élément de base des systèmes phonologiques, même s'il pose encore un certain nombre de problèmes non-résolus, à introduit l'aspect dynamique en phonologie, par la nature spatio-temporelle des gestes. Le modèle offre des solutions élégantes et naturelles à des problèmes qui se résolvaient autrefois par la formulation de règles 'ad hoc'.

Deux problèmes importants auxquels le modèle est confronté sont la détermination du paramètre k dans l'équation qui donne l'état de la variable du conduit (la raideur du ressort) et les moyens d'observation des gestes. Il paraît en effet assez difficile de déterminer objectivement la raideur du ressort pour un geste particulier. Par exemple, comment déterminer la valeur de k pour un articulateur tel que la pointe de la langue, afin de contraster la position du ressort pour une voyelle antérieure fermée, une fricative alvéolaire, une occlusive alvéolaire et une approximante ([i, s, z, t, d, j]). Les moyens d'observation des gestes sont assez divers et fonction des articulateurs observés. La cinéradiographie, n'est à présent plus utilisée pour des raisons qu'il n'est pas possible de mettre en cause. Les progrès de l'imagerie par résonance magnétique (IRM) permettent d'envisager de trouver de bonnes solutions de remplacement, mais quelques problèmes importants doivent encore être solutionnés, le nombre d'images que l'on peut obtenir dans un délai de temps le plus court possible et surtout la synchronisation du son et de l'image. Ceci fait que les gestes doivent être observés et mesurés par un ensemble de techniques qui sont assez différentes et qui ne sont par toujours mesurables en parfaite synchronisation. Les gestes peuvent être observés et mesurés directement en imagerie par résonance magnétique (IRM), en éléctromyographie (EMG) et en électropalatographie (EPG); il peuvent aussi être inférés à partir de mesures aérodynamiques et acoustiques.

3.2. *Auto-organisation*

Lorsqu'on étudie les systèmes phonologiques dans la perspective de l'évolution du langage, plusieurs questions importantes viennent immédiatement à l'esprit. Quand et comment les sytèmes phonologiques sont-ils apparus? Existe-t-il des sytèmes primitifs au sens où on pourrait rencontrer des systèmes reflétant un state ancien de l'évolution du langage? Comment les sytèmes se structurent-ils dans les sociétés? Comment évoluent-ils? La réponse à ces questions est fondamentale pour comprendre les mécanismes qui gèrent le fonctionnement normal des systèmes phonologiques des langues humaines. De nombreuses hypothèses ont été proposées pour expliquer l'émergence des systèmes phonologiques au cours de l'évolution, mais beaucoup de points restent encore obscurs. Nous ne disposons pas, par exemple, d'une explication satisfaisante pour expliquer la discontinuité qui existe entre le système de communication des primates supérieurs et celui de l'homme. Il n'est pas possible de montrer, en se fondant sur les données linguistiques dont nous disposons, qu'il a existé des sytèmes 'primitifs' dont on pourrait retrouver la trace dans les langues humaines ou dans leur histoire connue. Il est probable, que cela soit dû au fait, que le laps de temps pour lequel on dispose de données sur le langage humain est trop court, pour que des changements significatifs explicables en termes de modifications génétiques puissent être mis en évidence. L'écart qui existe entre les hommes modernes et les primates paraît si important, en termes de communication, qu'il semble difficile de montrer comment la parole ait pu évoluer des primates à l'homme moderne, malgré le lien évident qui existe entre les deux espèces au niveau de l'évolution.

[Lin83], [Ken81], [Ken84], [The89], [The91], [Dem96c] ont proposé de répondre à certaines de ces questions en montrant que le système de parole se fonde sur des principes auto-organisés. Ces principes fonctionnent à plusieurs niveaux, en ontogenèse lorsque le système se constitue chez l'individu et pour structurer le système au niveau du groupe.

Un concept important des systèmes dynamiques, qui est mis en en évidence par le paradigme d'auto-organisation, est celui de propriété émergente. Une propriété émergente

apparaît naturellement grâce à la dynamique du système, par la combinaison d'éléments préexistants qui forment des structures coopératives. Par exemple, la coopération des mouvements articulatoires est une proporiété émergente dans le sens où la coordination est permise par la dynamique de la coopération des mouvements. Ce concept d'émergence en phonologie est un mécanisme spontané qui complète les données dont dispose l'enfant pour développer son système phonologique. Lindblom [Lin92] propose de considérer les unités phonologiques comme émergentes. A la suite de [Stu87], Lindblom considère que l'information génétique seule est insuffisante pour rendre compte de l'acquisition du langage. Cet argument combiné à la pauvreté du stimulus dont dispose l'enfant pour acquérir le langage fait penser que l'interaction de ces 2 éléments ne produit pas seulement leur somme, mais quelque chose de plus. Cette interaction qui dépasse les conditions initiales, génère des structures émergentes dont la complexité n'est pas explicitement formulée, mais qui apparaissent par auto-organisation, comme une conséquence automatique de l'interaction entre les éléments. L'approche auto-organisée est utile pour trouver des explications au développement phonologique et à la génération des sytèmes de sons. Pour [Lin92] les formes phonétiques sont des adaptations aux contraintes de la production et de la perception aux facteurs spécifiques de la langue et de l'enfant. Dans la conception de Lindblom, les gestes phonétiques sont émergents et ne sont pas des primitifs de la théorie. Lindblom [Lin86] a montré, par des méthodes déductives, comment on peut prédire avec une assez bonne précision, la structure des systèmes vocaliques. Dans ce cas, on considère que la structure émerge de l'interaction entre les éléments du système, sans que la structure ne soit spécifiée au préalable. Des approches complémentaires et similaires ont été développées par [Car96] et [Sch97a], [Sch97b]. Un autre point développé par [Lin88] à propos des systèmes sonores des langues concerne leur degré de complexité. Lindblom et Maddieson [Lin88] ont donné une mesure de 'l'élaboration articulatoire' en proposant les critères, *basique*, *élaboré* et *complexe* pour rendre compte de la complexification croissante des phonèmes. Un segment basique n'a aucune élaboration, un segment élaboré comporte un seul degré d'élaboration et un segment complexe a deux ou plusieurs degrés d'élaborations. Selon ce principe, une rétroflexe [ɖ] est élaborée, une rétroflexe soufflée [ɖ̪] est complexe et une alvéolaire [d̪] est basique.

Un dernier point mérite d'être brièvement considéré dans cette perspective. Il s'agit de la structuration des systèmes phonologiques au niveau du groupe. L'auto-organisation collective [Den89], [Thé97] peut fournir quelques principes intéressants pour répondre à ces questions. Ce qu'il faut expliquer dans ce cas, c'est comment certaines variations non-systématiques du signal de parole sont amplifiées pour être intégrées dans le système phonologique partagé par une collectivité d'individus. Une explication possible à ce phénomène vient de la mise en oeuvre de la capacité mimétique dans l'amplification des variations. Donald [Don91] a montré que la capacité mimétique est une des adaptations motrices majeures qui a permis le développement du langage et de la phonologie. On peut penser que combinée à des variations phonétiques favorisées par des environnements sociolinguistiques déterminés, cette capacité mimétique joue un rôle important dans l'amplification des variations. Dans ce cas, il s'agit d'une boucle de rétroaction positive dont une conséquences importantes au niveau collectif est de mémoriser ces variations et de permettre en dernière instance leur intégration dans un système phonologique. Cette partie concernant l'auto-organisation est encore assez spéculative et peu explorée, mais elle offre des perspectives prometteuses pour comprendre les phénomènes de structuration collective en phonologie.

3.3 Diachronie

Systèmes vocaliques

Un domaine privilégié pour l'étude du comportement dynamique des sytèmes phonologiques est bien entendu la diachronie. Pour poursuivre brièvement la discussion sur la typologie des systèmes vocaliques et leur prédiction, qui est esquissée plus haut, il est utile d'examiner comment les systèmes vocaliques se comportent en diachronie. Un cas particulièrement intéressant se rencontre dans un sous-groupe de langues du phyllum Nil-Sahara des langues africaines. Il s'agit de l'évolution des systèmes vocaliques des langues moru-mangbetu dont l'inventaire passe de 10 à 6 voyelles et dont tous les stades intermédiaires existent encore [Dem88] et [Tuc56]. L'évolution des systèmes dans ce petit groupe de 25 langues est éclairante parce qu'elle permet de voir si la réduction du nombre de voyelles dans les inventaires de chaque langue, suit les mêmes tendances que celles qui sont proposées par les méthodes déductives pour expliquer la structuration des systèmes.

A l'exception d'un système à 6 voyelles et de 2 systèmes à 7 voyelles, les inventaires à 10, 9, 8 et 7 voyelles sont ceux que l'on trouve le plus souvent dans les systèmes à 10, 9, 8 et 7 voyelles de la base de données UPSID [Mad84] et dans la typologie de [Val84]. Vallée [Val94] souligne que lorsque les systèmes vocaliques ont plus de 5 voyelles, on trouve toujours une voyelle centrale dans le système, ce qui est le cas pour le seul système à 10 voyelles /i, ɪ, e, ɛ, a, ə, ɔ, o, ʊ, u/ (moru) et celui à 8 voyelles /i, ɪ, ɛ, a, ɔ, o, ɵ, u/ (lendu), mais pas pour le système à 6 voyelles /i, ɪ, ɛ, a, ɔ, u/ (kaliko) bien que ce dernier ait une voyelle centrale, mais qui n'a pas (ou plus) de statut phonologique. Les systèmes à 9 voyelles /i, ɪ, e, ɛ, a, ɔ, o, ʊ, u/ (les langues mangbetu-asua et mangbutu-efe, le madi et le lulubo) ont un nombre équivalent de voyelles antérieures et postérieures. Les inventaires à 7 voyelles sont de deux types /i, ɪ, ɛ, a, ɔ, ʊ, u/ (lugbara) et /i, ɪ, ɛ, ə, a, ɔ, u/ (logo et avokaya). Les systèmes à 7 voyelles du

logo et de l'avokaya et le système à 6 voyelles du kaliko sont des systèmes peu fréquents dans les langues du monde, remarquons cependant que ces langues se distinguent par le maintien d'une opposition entre les 2 voyelles antérieures fermées /i et ɪ /.

Le passage de 10 à 9 voyelles voit disparaître la voyelle centrale [ə] ; le passage de 9 à 8 voyelles montre la fusion des timbres [e et ɪ] et l'apparition d'une voyelle centrale [ɵ ou ɯ] qui a de nombreux allophones en Lendu ([i, ʉ]), [Dim86]. Le passage à 7 voyelles voit disparaître, soit la voyelle centrale [ɵ], soit la fusion des timbres [o et ɔ]; le passage de 7 à 6 voyelles fusionne [ə et a]. On peut résumer cette évolution de la manière suivante (une voyelle qui disparaît est représentée par le symbole - afin d'éviter toute confusion avec les voyelles [o] et [ø]) :

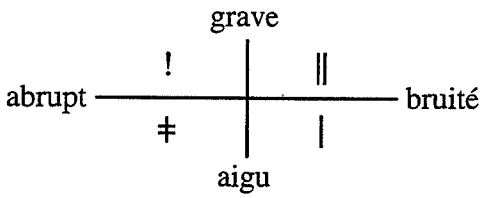
10 > 9	ə > -
9 > 8	e > ɪ et ʊ > ɵ
8 > 7	ɵ > - ou o > ɔ
7 > 6	ə > a

Ces changements indiquent que les facteurs qui déterminent la dispersion et la discriminabilité des voyelles jouent aussi un rôle déterminant en diachronie pour expliquer la structure et la stabilité des systèmes. On constate par exemple que les voyelles centrales qui sont perceptuellement moins proéminentes ont tendance à disparaître plus vite que les voyelles périphériques. Globalement l'évolution du timbre des voyelles des langues moru-mangbetu semble plus tenir compte des contraintes perceptuelles que d'environements phonétiques particuliers.

Consonnes et clicks

Les données diachroniques offent de très nombreux exemples pour illustrer la dynamique des systèmes phonologiques. Un magnifique cas est donné par Traill et Vossen [Tra97] qui montrent comment, dans les langues khoisan, le principe de 'facilité d'articulation' explique la passage de clicks 'fortis' aux clicks affaiblis et enfin à la perte des clicks.

Traill et Vossen [Tra95] et [Tra97] montrent qu'il existe deux traits acoustiques binaires indépendants pour caractériser l'influx d'air des clicks: l'un est comme une impulsion avec un rapide amortissement opposée à un influx bruité avec un amortissement lent; l'autre se défini par une accentuation des hautes ou des basses fréquences du spectre. Le premier trait est défini comme 'abrupt' et le second comme 'grave'. La représentation des clicks dans cet espace acoustique proposée par Traill et Vossen est donnée comme suit:



Pour Traill et Vossen, l'affaiblissement et la perte des clicks sont simplement la réponse à des pressions de la coarticulation du contexte phonétique, à l'encontre duquel les clicks abrupts requièrent une distorsion extrême des trajectoires de la langue. Affaiblir ces clicks implique un 'undershoot' articulatoire mais comme le résultat est un click bruité, plus faible et diffus, il y a un coût acoustique. Ce coût peut être vu comme la réduction des clicks dans l'espace acoustique, les clicks abrupts se déplaçant vers le centre de leurs quadrants, ce qui a pour effet d'atténuer la distinction auditive des clicks. Traill et Vossen suggèrent, en outre, que ce compromis pourrait expliquer pourquoi l'étape suivante de 'l'undershoot' articulatoire est implémentée. Cette étape donne lieu à la perte du click, sous la forme d'un non-click acoustiquement similaire et dont l'effet est de rétablir la proéminence parmi les contrastes [Tra97]. Ce point illustre parfaitement la notion de contrastivité qui joue un rôle essentiel dans la dynamique des systèmes phonologiques, à la fois pour maintenir la stabilité des systèmes et pour les faire évoluer. Le passage du click alvéolaire au click latéral: [! > ‖] est caratéristique du phénomène d'affaiblissement. La perte des clicks se caractérise par les changements suivants: [! > k] et [ǂ > c]. L'exemple des changements phonétiques en khoisan montre clairement que pour expliquer le comportement dynamique des sytèmes phonologiques et de certains de leurs composants, il est nécessaire de tenir compte de tous les facteurs qui sont impliqués en production et en perception de la parole. Les changements phonétiques qui concernent les clicks sont de très bons exemples, d'une part de la régularité des changements, et d'autre part du fait que pour expliquer l'évolution des systèmes il est absolument nécessaire de recueillir le plus de données comparatives et diachroniques possibles. Ces données linguistiques combinées à des observations expérimentales permettent non seulement de comprendre l'évolution des systèmes phonologiques, mais aussi quels sont les paramètres importants qui gèrent leur fonctionnement. L'explication des changements phonétiques en khoisan montre que pour expliquer leur dynamique, des données tant aérodynamiques, qu'articulatoires et acoustiques sont nécessaires.

4. Aspects expérimentaux

Parmi les principaux problèmes qu'une approche expérimentale et dynamique de la phonologie doit résoudre, figurent la recherche de preuves expérimentales de l'interaction entre les différents niveaux du système de

production de la parole, ainsi que la compréhension de la manière dont les composants qui forment les niveaux sont contrôlés et régulés. L'étude de la dynamique des sytèmes phonologiques peut être abordée de manières différentes du point de vue expérimental. Le modèle de phonologie articulatoire met l'accent sur les gestes, il est donc nécessaire de trouver des moyens pour caractériser la trajectoire et la temporalité des articulateurs (les variables du conduit) qui sont impliqués dans chaque geste. Ceci peut être réalisé de différentes manières, soit en mesurant directement la trajectoire des articulateurs en cinéradiographie, ou de certains articulateurs au moyen d'un 'move-track' ou d'un articulatographe. D'autres façons d'aborder les problèmes de la phonologie du point de vue expérimental sont proposées par [War96] qui fournit plusieurs exemples de régulation de l'aérodynamique en production de la parole. Cet aspect de l'étude de la parole est crucial si l'on veut comprendre comment le système de production et les systèmes phonologiques sont régulés.

Les moyens d'investigation dont on dispose pour faire de l'expérimentation sont assez nombreux, mais ils sont souvent limités par le cararctère envahissant qu'ils peuvent revêtir. Les principales méthodes expérimentales qui permettent de traiter de problèmes concernant la production de la parole sont décrites dans [Sto96] et [Sto97]. On pourrait aussi ajouter les nombreux paradigmes qui permettent d'étudier les problèmes de perception (voir par exemple [Gro97] pour un inventaire des principaux paradigmes liés à la reconnaissnce des mots), mais cet aspect de la question ne sera pas abordé ici.

Afin d'illustrer le problème de la dynamique articulatoire de manière élémentaire, on peut brièvement s'attarder à la description de segments dynamiques comme les flaps labio-dentaux [v̆] qui se rencontrent en mangbetu, une langue de la famille soudan central parlée au Congo [Dem96b]. Ces sons, dont la durée totale du geste est comparable à celle des autres consonnes de la langue, sont produits en rétractant la lèvre inférieure à l'intérieur de la bouche de 100 à 120 ms, puis en la projettant vers l'avant en 20 ms, de sorte à réaliser une brève occlusion (également de l'ordre de 20 ms) lorsque la lèvre inférieure heurte les dents de la mâchoire supérieure. On obtient ainsi une articulation dynamique, par le mouvement quasi ininterrompu des articulateurs impliqués. La figure 3 illustre schématiquement, à partir d'images vidéo prises de profil, la position des lèvres au début du flap et la position de la lèvre inférieure au moment de sa rétraction maximale.

La mesure des paramètres aérodynamiques et acoustiques permet ensuite d'inférer, notamment, la temporalité du mouvement des articulateurs [Dem96]. Ce type d'articulation montre l'interaction qui existe entre certaines des variables qui permettent la réalisation d'un flap labio-dental ainsi que l'aspect dynamique de ce segment.

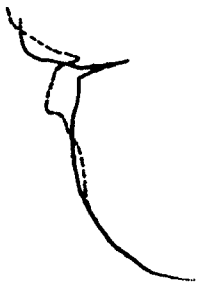


Figure 3: Position schématique des lèvres au début (ligne pointillée) et au moment de la rétraction maximale (ligne continue) de la lèvre inférieure pour un flap labio-dental.

Le fonctionnement normal du système de parole et d'un système phonologique est typique d'une régulation homéostatique dont les principes de base sont (i): la régulation en vue de maintenir la stabilité du système et (ii) le contrôle des mécanismes pour obtenir des conditions relativement stables; le but d'une régulation homéostatique étant de maintenir un paramètre du système à un niveau acceptable, un peu à la manière d'un termostat qui régule la température d'une pièce (voir aussi [Tho95], pour plus de détails). (War96] fait une remarque importante lorsqu'il affirme qu'il est nécessaire de distinguer les termes de *contrôle* et de *régulation*. On peut ainsi désigner le contrôle comme le processus de gestion et régulation peut être défini comme les contraintes qui ajustent le degré de production des éléments à l'état du système et aux différentes variables environnementales.

L'étude systématique d'un système de régulation nécessite des expériences pour contrôler et identifier les mécanismes de contrôle. Pour illustrer quelques aspects du contrôle et de la régulation dans les composants ou entre les composants de la production de la parole, trois problèmes vont à présent être abordés; celui du contrôle et de la compensation des mouvements articulatoires; celui de la régulation de la pression à l'intérieur du conduit vocal et celui de l'information proprioceptive que nous possédons du composant aérodynamique.

4.1 Contrôle et compensation

Des expériences consistant à produire des ouvertures inhabituelles de la mâchoire ont été conduites pour étudier les réponses aérodynamiques à des modifications soudaines dans le conduit vocal. Ces expériences forcent les locuteurs à des maneuvres compensatoires qui peuvent être évaluées quantitativement [War80], [War81], [War84]. Les résultats montrent un comportement remarquablement adaptatif des locuteurs aux obstacles présentés [War96]. D'autres expériences conduites par [Lin79], [Lub79] et [Fow80] afin d'étudier les compensations du point de vue acoustique montrent que lorsqu'on fait obstacle aux mouvements de la mâchoire en demandant aux locuteurs de

mordre une allumette, on obtient aussi des compensations immédiates. Ces compensations peuvent aussi être étudiées, du point de vue articulatoire, au moyen de l'imagerie par résonance magnétique dynamique [Dem97b]. Ces mouvements articulatoires de la langue sont différents lorsqu'il est demandé à des locuteurs de mordre une petite bouteille en plastique contenant de l'eau, quand ils produisent une série ininterrompue de voyelles [uiu]. Ces compensations sont illustrées à la figure 4 qui montre comment la masse de la langue se déplace lorsque la mâchoire inférieure est bloquée, par rapport à son mouvement normal.

Bien que difficilement quantifiables par cette méthode à l'heure actuelle, ces compensations sont intéressantes parce qu'elles impliquent la mise en oeuvre rapide et contrôlée de muscles différents. Ces données confirment les affirmations de [Dea97] qui montre qu'il existe un contrôle cortical direct sur les muscles de la langue. Ceci signifie donc que la notion de contrôle implique que le cerveau reçoit cette information, la traite et dirige le contrôle de l'activité des muscles de la langue.

Dans le fonctionnement normal de la parole, il est aussi possible d'observer des différences intéressantes dans la temporalité des mouvements de certains articulateurs, au sein d'une même langue. On constate ainsi que la différence entre les voyelles nasales du français méridional et celles des dialectes du nord tient à une ouverture plus rapide du voile dans les parlers du nord, ce qui a pour effet de nasaliser complètement la voyelle. Dans les parlers méridionaux, le voile s'ouvre plus tadivement et achève son geste par une occlusion vélaire. La figure 5 montre la différence de position du voile au début et à la fin d'une voyelle nasale dans les deux parlers. Ceci indique que la mesure du déplacement et de l'aspect temporel des gestes est fondamentale si l'on veut comprendre la dynamique des systèmes phonologiques dans lesquels ils sont intégrés.

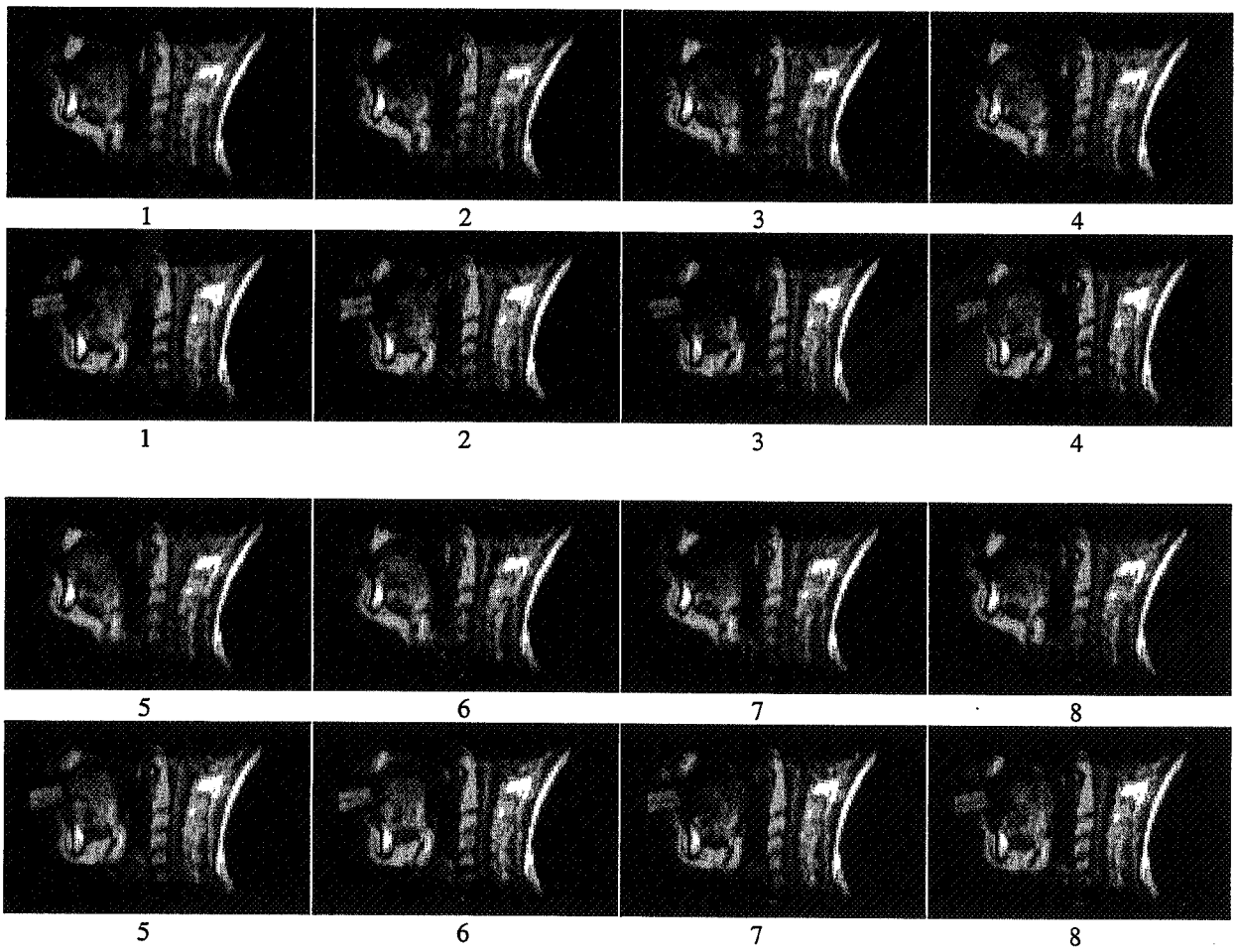


Figure 4: Comparaison des mouvements de la langue pour un locuteur exécutant la séquence [uiu]. La première montre le mouvement sans obstacle. Dans la deuxième série la mâchoire est bloquée par une bouteille d'eau tenue entre les dents par le sujet. L'intervalle de temps entre deux images est de 250 ms.

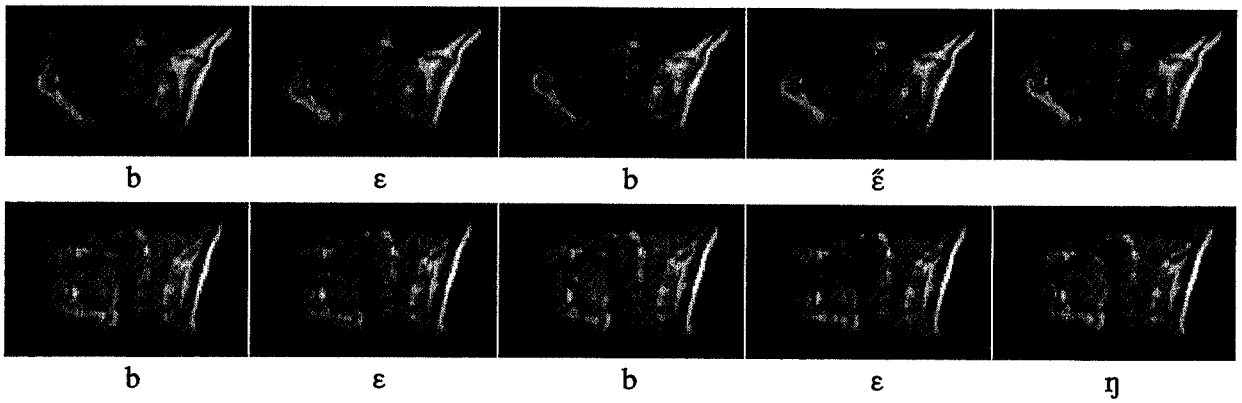


Figure 5: Comparaison des mouvements du voile du palais exécuté pendant les logatomes [bɛbɛ̃] et [bɛbɛŋ]. La première série d'image est produite par un locuteur belge, la seconde par un locuteur méridional. L'intervalle de temps entre deux images est de 250 ms.

4.2 Regulation et pression

Les pressions sous-glottiques (psg) et supra-glottiques (pio) jouent un rôle important dans la régulation de l'aspect aérodynamique de la parole. La psg qui est la source d'énergie primaire pour la parole est gardée relativement constante dans la parole normale [Lad71], [Oha76]. Le maintien de la psg ne dépend pas du seul mouvement des forces élastiques. Les ouvertures et les fermetures du conduit vocal nécessitent des compensations dans des délais très brefs si la pression doit être maintenue plus ou moins constante. Les mesures de pio varient beaucoup plus que la psg et sont fonction de nombreux facteurs. Il est évident que si la pression est régulée dans le système de parole, c'est la psg qui sera maintenue plus ou moins constante. [War82] suggère que les résistances des structures de la parole sont contrôlées de manière synchronisée et analogue au système respiratoire. Les figures 6 et 7 comparent des mesures de psg, de pio et le signal acoustique pour les consonnes et les voyelles [n, R, p, b s, z]. Chaque consonne est réalisée dans un contexte CVCV où la voyelle est [a]. Les données de la figure 7 permettent de constater que la psg est maintenue relativement constante, malgré des fluctuations qui sont dues à des modifications de la résistance à la glotte, lorsqu'on prononce des occlusives, des fricatives, des nasales ou des trilles. Sur cette figure, on peut aussi voir que les fluctuations de la psg sont nettement plus importantes dans les séquences qui impliquent des consonnes sourdes [p, s].

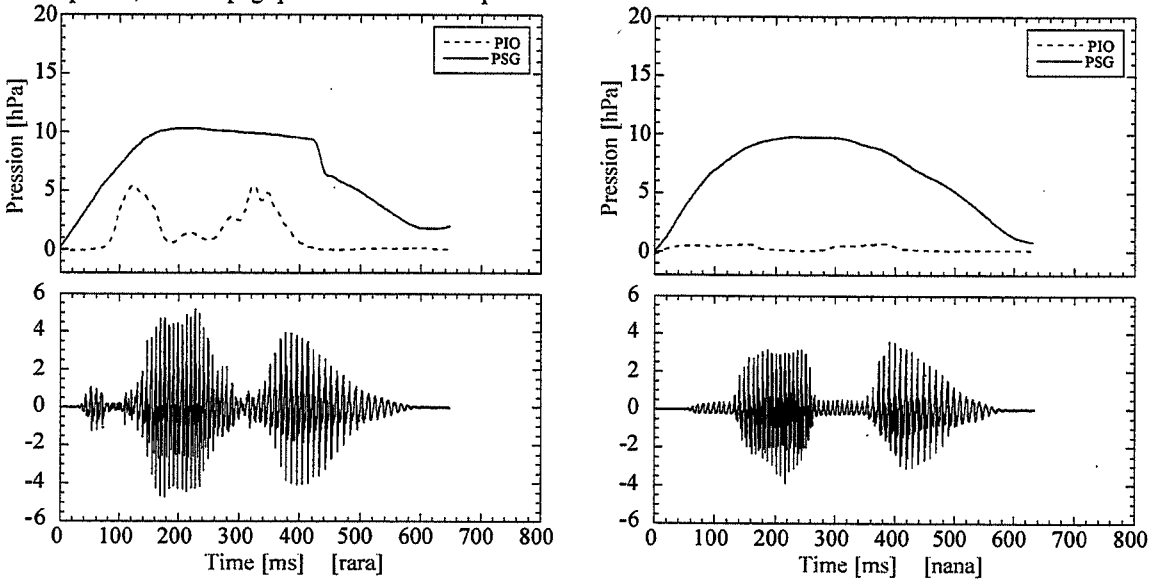


Figure 6: Mesures de pression sous-glottique, de pression intra-orale et du signal de parole dans des logatomes CVCV ([rara] et [nana]) prononcés par un locuteur du français.

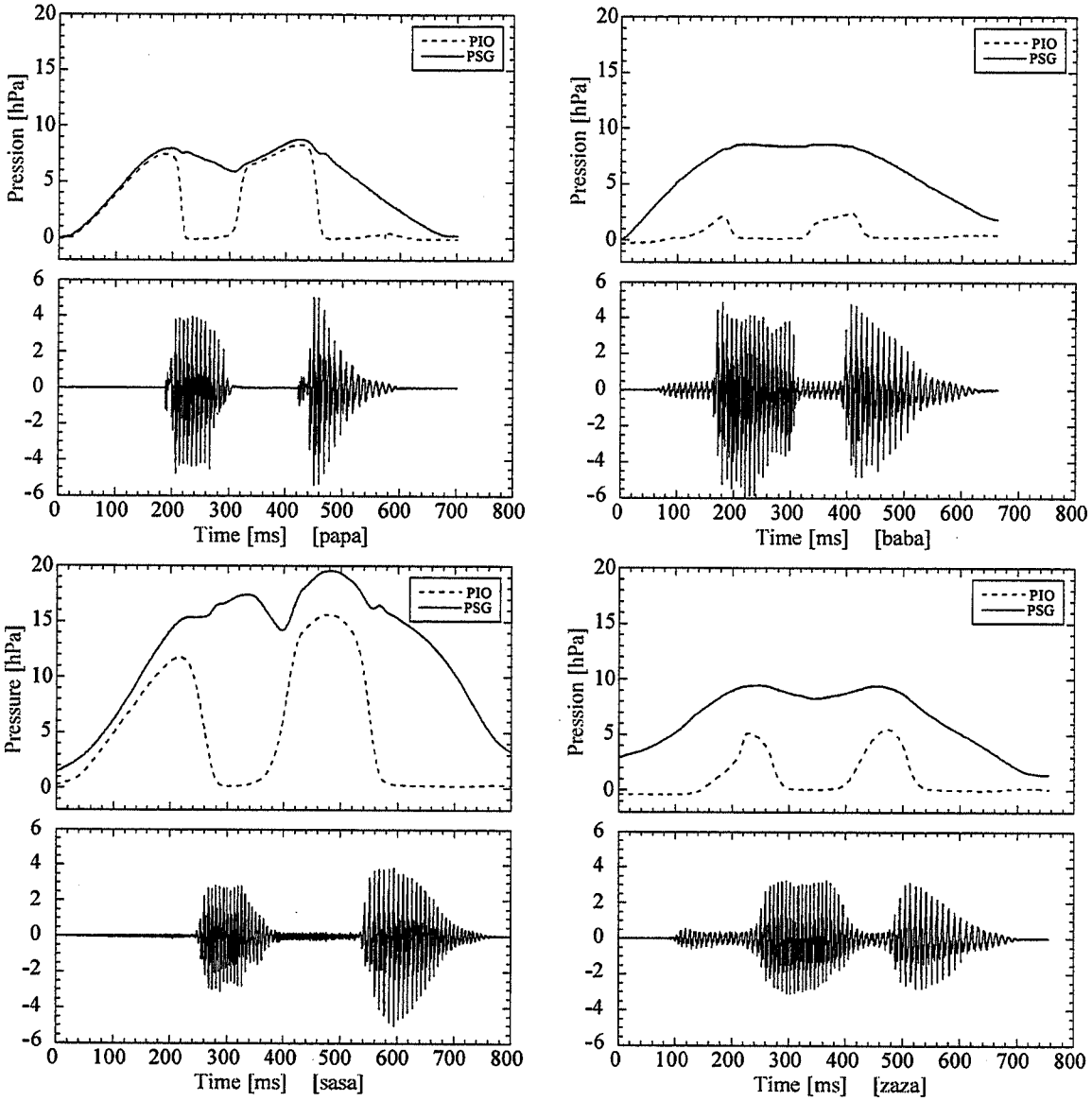


Figure 7: Mesures de pression sous-glottique, de pression intra-orale et du signal de parole dans des logatomes CVCV ([papa], [baba], [sasa] et [zaza]) prononcés par un locuteur du français.

Un autre cas de régulation de la pression est illustré par [War81] qui montrent que la relation entre la pression intra-orale et le débit d'air buccal est similaire pour les consonnes fricatives. Les différences de grandeur de la constriction ont peu d'effet sur la pression intraorale pendant la production des fricatives parce que le débit d'air buccal covarie avec la grandeur de l'ouverture. Ces exemples illustrent l'importance du composant aérodynamique dans la régulation du système de parole (voir aussi [Moo93]).

Un dernier cas intéressant qui concerne la pression se rencontre dans des consonnes à articulation double, les labio-vélaires ou les labio-uvulaires [Dem96a], [Dem97a]. Ces consonnes se rencontrent dans un nombre important de langues africaines [Lad96]. Les labio-uvulaires [q͡p] se rencontrent en mamvu, en lese et en efe [Vör74]. Ces consonnes sont intéressantes parce que la partie uvulaire de l'articulation est sourde et que la partie bilabiale est implosive et sonore. Rappellons qu'il s'agit ici d'un seul phonème et non d'une séquence de deux consonnes. Il faut donc expliquer comment ce type d'articulation est réalisée et contrôlée, puisqu'il s'agit d'un élément du système phonologique de ces langues. Dès le moment où ce type de consonne est intégrée au système phonologique de la langue et n'est pas ou plus une simple variante, un allophone, d'un autre phonème, il faut un contrôle sur la séquence d'événements articulatoires qui la génèrent. Dans ce cas, on peut penser qu'il existe aussi une régulation sur la pression à l'intérieur des deux cavités, mais ceci reste à démontrer. La question de savoir si notre système de détection est sensible à la pression dans le conduit vocal reste ouverte, mais il existe de nombreux indices qui

permettent de penser que c'est le cas [Mal66], [Mal 70], [Mül80] et [Wyk81].

Les parties sourdes et sonores des consonnes labio-uvulaires du lese ne sont pas réalisées avec une ouverture et une fermeture des cordes vocales au cours de la même articulation, mais par un simple changement des conditions aérodynamiques dans le conduit vocal. La consonne commence par une occlusion uvulaire qui est suivie d'un rapide équilibre des pressions sous-glottiques et supra-glottiques. Cet équilibre des pressions met fin à la vibration des cordes vocales. Ensuite, la fermeture labiale se réalise, suivie par un mouvement d'abaissement de la mâchoire inférieure et d'abaissement et de recul de la langue qui est aisément détectable par la mesure de pression buccale qui devient négative, suite à l'augmentation progressive de la cavité buccale. Le recul et l'abaissement de la langue ont pour effet de relâcher l'occlusion uvulaire ce qui provoque un appel d'air à l'intérieur de la cavité buccale, et qui a pour conséquence de remettre les cordes vocales en vibration. Le mouvement d'abaissement du larynx se poursuit quelques ms et est suivi du relâchement de l'occlusion labiale. La figure 8 montre le signal et les mesures de pression prises dans le pharynx et dans la cavité buccale pendant le mot hagɓoqɓu 'crapaud' en lese. Notons aussi que ce mot contient une autre consonne double, la labio-vélaire [gɓ]. La figure 9 montre la séquence des mouvements articulatoires qui sont réalisés pour produire cette consonne.

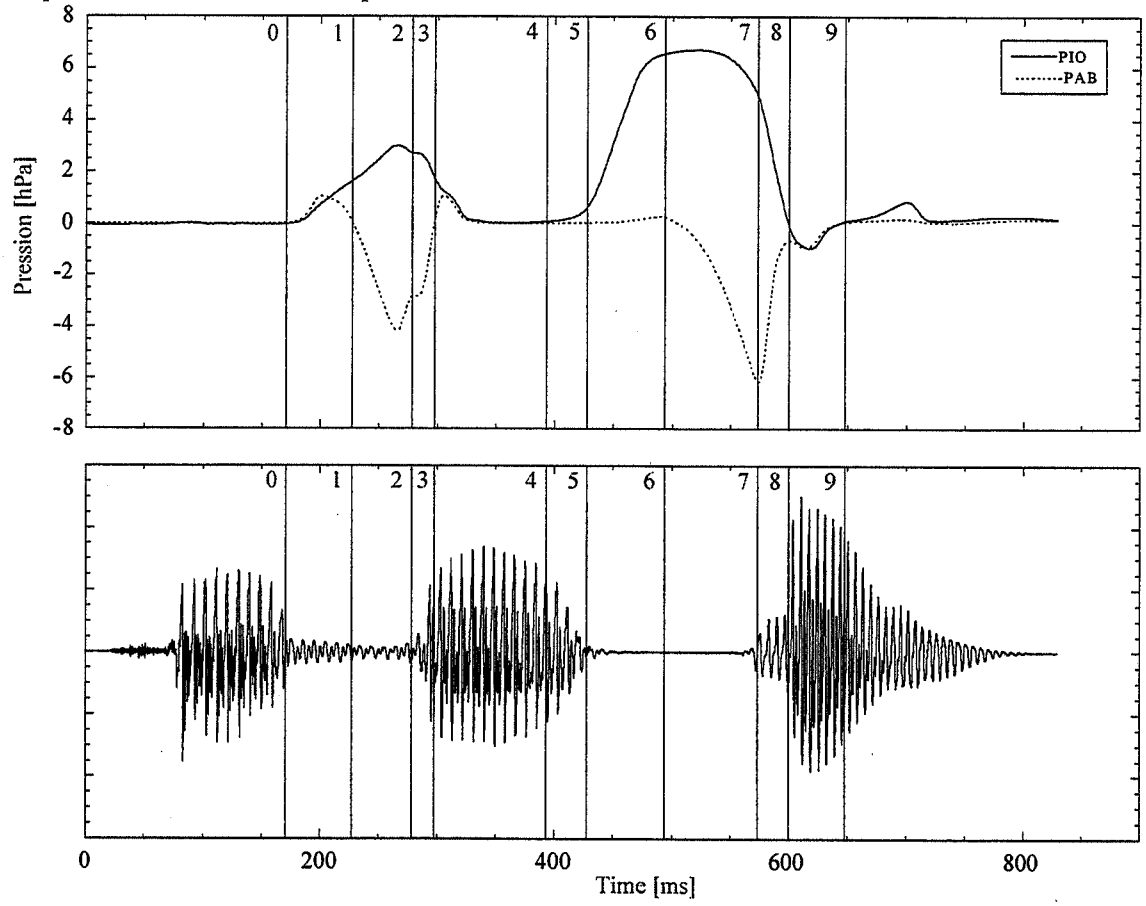


Figure 8. Pression intra-orale (pio), pression d'air buccale (pab) et signal de parole dans le mot hagɓoqɓu 'crapaud' en lese. Les curseurs 0 à 3 concernent [gɓ] et 4 à 9 [qɓ]. 0) fermeture labiale et vélaire 1) abaissement de la masse de la langue et de la mâchoire 2) effet de l'abaissement du larynx sur la pression et le voisement 3) relâchement de la fermeture labiale 4) début de la fermeture labiale 5) arrêt du voisement 6) début de l'abaissemnt de la langue et du larynx 7) relâchement de la fermeture uvulaire 8) relâchement de la fermeture labiale 9) retour du larynx à la position initiale.

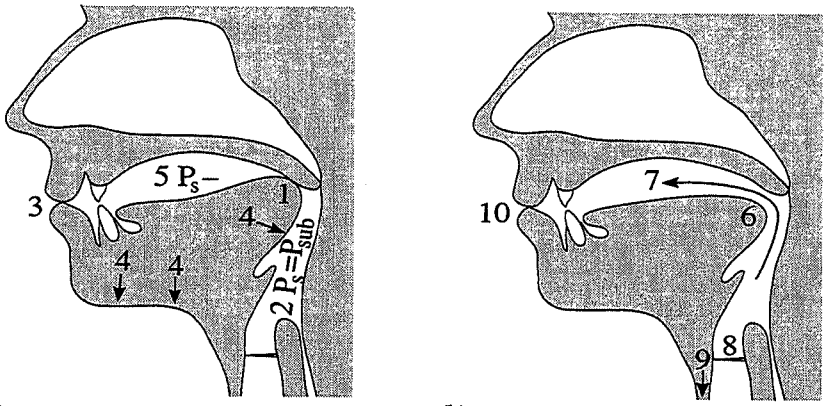


Figure 9a et 9b. Séquence d'événements montrant la production de la consonne labio-uvulaire [qɓ]. 1) fermeture uvulaire 2) équilibre des pressions sous-glottique et supra-glottique 3) fermeture des lèvres 4) abaissement de la masse de la langue et de la mâchoire 5) expansion de la cavité orale 6) relâchement de la fermeture uvulaire 7) appel d'air dans la cavité buccale 8) redémarrage du voisement 9) abaissement du larynx 10) relâchement de la fermeture labiale.

4.3 Information proprioceptive et sensorielle

Comme tout système de régulation, la parole comprend une relation entre l'entrée, le traitement et la sortie. Le problème des capteurs sensoriaux a été mentionné à propos des labio-uvulaires, mais comme cela a été souligné plus haut, il faut encore démontrer leur existence, trouver leur localisation et mesurer leur sensibilité. Ceci est important pour comprendre les facteurs de contrôle qui opèrent et aussi obtenir des informations sur les rétrocations entre les composants de la production de la parole, par exemple entre les composants articulatoires et aérodynamiques.

L'expérience suivante permet de comprendre et de mesurer les relations entre les gestes articulatoires et les facteurs aérodynamiques. Un sujet qui a un palais artificiel en bouche doit produire une série de consonnes fricatives dans des suites de mots français, dont un se termine par une fricative et un autre commence par une fricative identique à la première. Les suites de mots étaient les suivantes:

'C'est une nasse serrée', 'une niche chocolat', 'une glace sorbet', 'un étrange jouet'. La pression intra-orale est mesurée simultanément par un petit cathéter en plastique de 2mm de diamètre interne placé dans l'oro-pharynx, via la cavité nasale. Dans la seconde narine est placé un autre tube, en plastique rigide de 6mm diamètre interne, connecté à une pompe à air. Lorsque le sujet prononce les différentes fricatives, une éléctrovanne déclenche un appel d'air via la pompe à air, de manière à provoquer une brusque chute de pression à l'intérieur du conduit vocal. La pression intra-orale et les chutes de pression sont mesurées via la station de travail Physiologia [Tes90]. Les mouvements de compensations articulatoires sont mesurés sur l'EPG, également via la station Physiologia. Le moment du déclenchement de l'électrovanne est mesuré par le bruit quelle provoque sur le signal acoustique. L'ensemble de ces mesures sont synchronisées avec Physiologia et le logiciel Phonédit qui gère la station de travail.

Les résultats des mesures d'EPG indiquent clairement qu'il y a une compensation articulatoire à la réduction du débit provoquée par la brusque chute de pression. Cette compensation se fait par une réduction de la constriction. Cette modification est plus importante que la réduction qu'il est normalement possible d'observer vers la fin des fricatives alvéolaires [s, z] et palatales [ʃ, ʒ] du français. La modification de la constriction se fait bien dans la région alvéolaire, mais aussi sur l'ensemble de la région couverte par le palais artificiel. La figure 10 montre la mesure de la pio et des mesures d'EPG prises pendant la petite phrase 'une nasse serrée'. L'électrovanne a été déclenchée vers le milieu de la fricative alvéolaire [s:]. On peut observer la chute de la pio vers le milieu de la fricative et son rétablissement vers la fin de la consonne. Les mesures d'EPG ont été faites avant le déclenchement de l'éléctrovanne, (1) lorsque la constriction était stabilisée et (2) au niveau le plus bas de la la pio après l'ouverture. La comparaison des deux ensembles de contacts linguaux-palataux montre qu'après le déclenchement de l'electrovanne, il y a un contact plus important sur l'ensemble du palais, ce qui indique une réduction compensatoire de la constriction. Cette phénomène est très rapide, le début de l'augmentation des contacts se situe en moyenne 55ms après le début de la chute de pression. Les mêmes effets ont été observés dans l'ensemble des données traitées. Ces résultats montrent qu'il y a bien une rétroaction du composant articulatoire sur le composant aérodynamique.

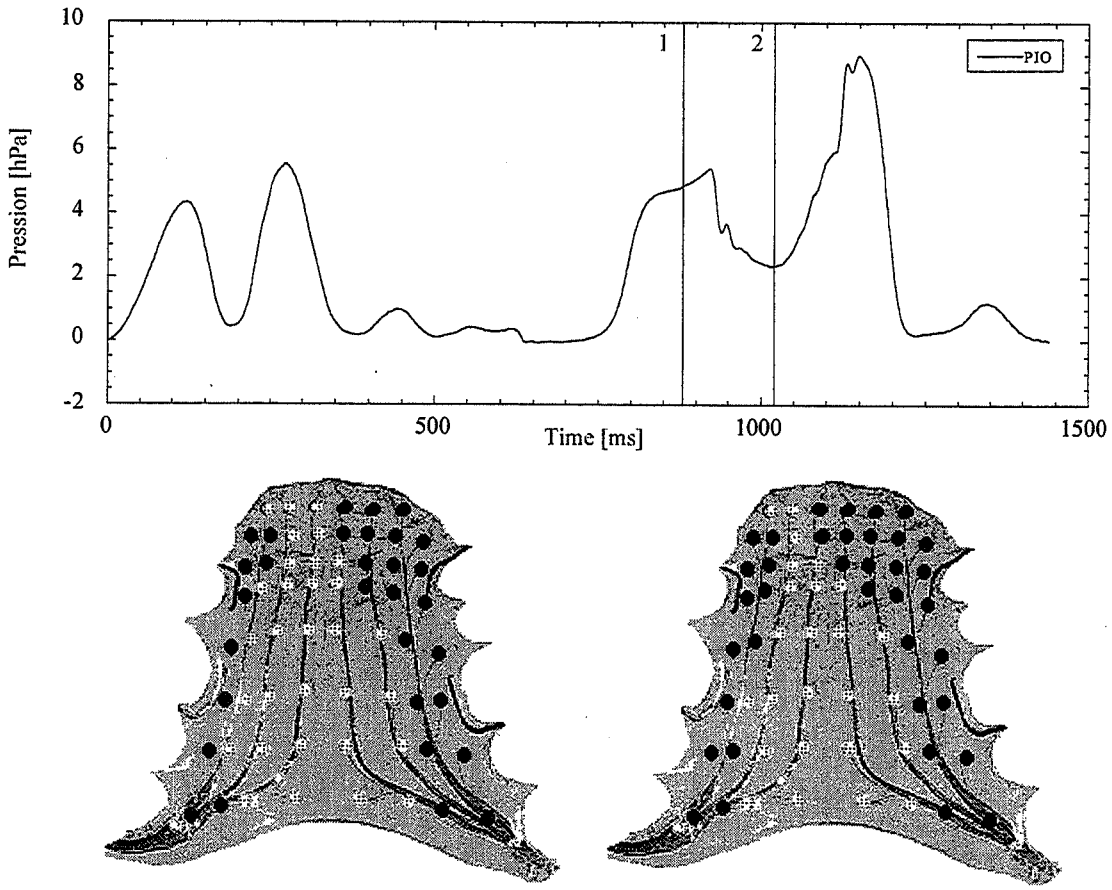


Figure 10: Pression intra-orale mesurée dans la phrase 'C'est une nasse serrée' ainsi que les points de contact EPG mesurés aux instants 1 et 2.

5. Aspects formels

Pour terminer, je voudrais brièvement aborder quelques aspects du traitement formel de la dynamique des systèmes phonologiques. Cette courte discussion se bornera à montrer l'intérêt qu'il y aurait à adopter un formalisme logique tel que celui qui a été proposé par Thomas et D'Ari [Tho90].

L'équation du mouvement, qui est proposée dans le modèle de Browman et Goldstein peut être utilisée pour identifier des attracteurs ou des solutions stables. Un aspect très important d'un attracteur est qu'il représente un point de convergence d'un système qui peut avoir eu différentes conditions initiales. Ceci signifie que quel que soit la position de départ du système, celui-ci peut s'installer dans certaines régions de stabilité. Plusieurs types d'attracteurs peuvent défnir des solutions stables. Un exemple d'attracteur dans la parole est le contact de la masse de la langue avec le palais dur. La langue peut avoir eu différentes positions de départ (différentes conditions initiales), elle devra en dernière instance atteindre le point contact prévu. L'intérêt d'un traitement dynamique de la production de la parole a été souligné à de nombreuses reprises, notamment par Kelso [Kel86] et par Saltzman [Sal89]. Thomas montre [Tho95] qu'à la différence du traitement en équations différentielles, un traitement employant un formalisme logique généralisé permet d'obtenir une vue analytique du système et non plus seulement une description de l'état du système. Les deux approches étant par ailleurs parfaitement complémentaires.

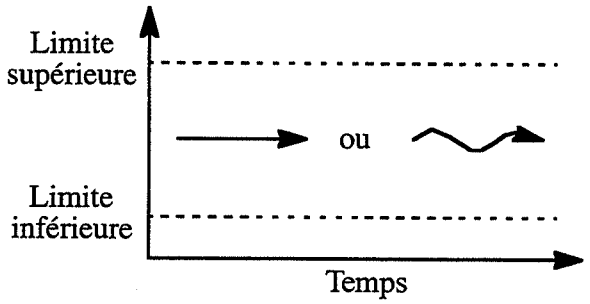
5.1 Régulations homéostatiques et différentiatives

Nous avons vu comment certains composants du système de production de la parole contribuent à sa régulation et donc au maintien de sa stabilité. Il s'agit par exemple, de la régulation de la pression sous-glottique par des ajustements de la glotte, ou encore de la rétroaction négative ou homéostatique de gestes articulatoires sur des modifications de pression dans le conduit vocal. Thomas montre [Tho90], [Tho95] qu'une régulation homéostatique peut être définie par une boucle de rétroaction négative, dans laquelle chaque élément exerce un contrôle négatif sur son propre développement via les autres éléments (pour autant qu'il y en ait) du système. Ce type de boucle de rétroaction caractérise le fonctionnement normal (la

stabilité) du système de la parole. Les boucles de rétroaction positives quand à elles jouent un rôle crucial pour expliquer les changements d'état d'un système phonologique.

A la suite de Thomas [Tho95] pour qui les boucles de rétroaction positives sont la modalité biologique de la multistationnarité, nous pouvons dire qu'elles reflètent les changements d'état que l'on peut observer dans les systèmes phonologiques. Une boucle positive rend compte de la mécanique des changements phonétiques et des changements d'état dans la structure des sytèmes. Les description de boucles positives constituent donc un élément essentiel pour la compréhension de la dynamique des systèmes phonologiques. La figure 11 illustre les concepts de régulation homéostatique et différentiative. La figure 12 montre une boucle de rétroaction positive, produit de l'interaction entre deux éléments (deux gestes) X et Y.

(a)



(b)

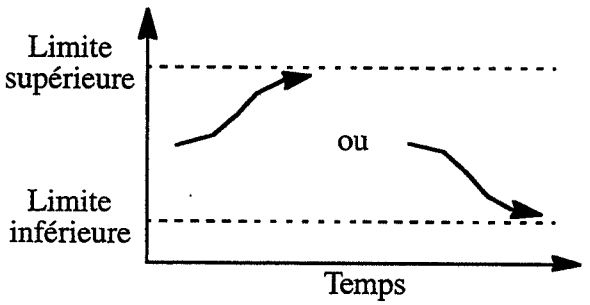


Figure 11: Illustration du comportement (a) d'une régulation homéostatique où on peut observer le maintien d'une valeur (stable ou oscillante) entre deux limites et (b) d'une régulation différentiative où on observe le changement d'état vers la limite inférieure ou supérieure de la valeur considérée.

Figure 12: Illustration d'une boucle de rétroaction positive.

5.2 *Implications de la régulation des systèmes phonologiques*

Un point important qu'il convient de souligner ici est que des systèmes qui ont des niveaux de complexité structurale très différents peuvent être extrêmement similaires du point de vue logique et donc, leur possibilités dynamiques être qualitativement similaires. Le choix des éléments ou des variables qui décrivent un système est crucial. Pour la parole, nous les définirons comme des gestes articulatoires et non par des traits pour des raisons qui ont été exposées à propos de la différence entre traits et gestes.

Considérons brièvement quelques notions du formalisme logique généralisé proposé par Thomas, [Tho90], [Tho91], [Sno93], [Tho95] pour illustrer un traitement possible d'éléments du système de la parole. La présentation est tirée de [Tho95]. Une variable logique x, dont la valeur décrit le niveau de l'élément, est associée à chaque élément pertinent du système. L'état présent du système est donc transcrit par un vecteur d'état. Dans le but de décrire l'évolution du système, à chaque élément on associe aussi un opérateur $X=f_x(x, y, z...)$ qui décrit les facteurs agissant sur l'évolution de x. Chaque combinaison des valeurs des variables (chaque état du système) a une valeur X. X est appelé l'image de x parce que sa valeur est la valeur vers laquelle x tend lorsque le système est dans l'état considéré. En d'autres termes, quand $X \neq x$, x a un ordre pour adopter la valeur de X. Dans cette perspective, l'évolution d'une variable dépend des simples règles suivantes:

(i) si X=x, la variable x est stable

(ii) si X>x, la variable x a un ordre pour augmenter sa valeur

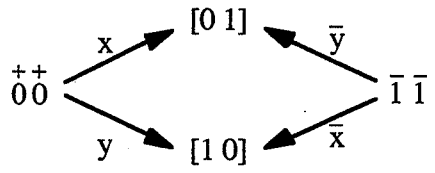
(iii) si X<x, la variable x a un ordre pour diminuer sa valeur

On pourrait ainsi dire dans le système de parole que x=1 veut dire que le produit du geste (un trait) est présent et X=1 que le geste est activé. Si le geste est activé, le produit est exécuté et sera présent après un délai de temps approprié. A n'importe quel instant de temps, le sytème peut être décrit par un vecteur d'état (x, y, z..) et son évolution par l'image (X, Y, Z, ...). Lorsque le vecteur d'état et l'image sont similaires, le système est dans un état logique stable. Lorsque l'état et l'image sont différents, il est essentiel de comprendre que l'image d'un état n'est habituellement pas son état suivant, simplement parce qu'habituellement une variable s'active ou se désactive en même temps. Si deux gestes ou plus sont activés par un signal commun, il faudra des délais de temps différents pour que les traits atteignent un niveau suffisant, de la même manière si les gestes sont désactivés.

Examinons le système $X=\bar{y}$, $Y=\bar{x}$ qui existe en phonologie: deux gestes qui se contredisent mutuellement dans le sens ou le trait de l'un empêche la réalisation de l'autre. La table d'état est:

$\overset{++}{x\ y}$	XY
$\overset{++}{0\ 0}$	1 0
[0 1]	0 1
[1 0]	1 0
$\overline{1}\ \overline{1}$	0 0

Les + et les - suscrits sur les éléments du vecteur d'état sont utilisés quand une variable et son image ont différentes valeurs logiques; xy/XY=00/11 peuvent être écrits de manière plus compacte, $\overset{++}{0\ 0}$. On peut constater que pour les états 01 et 10 l'état et l'image sont identiques. Dans les états $\overset{++}{0\ 0}$ et $\overline{1}\ \overline{1}$ il y a deux ordres. Dans l'état $\overset{++}{0\ 0}$ il y a un ordre pour activer x de 0 à 1 et un ordre pour activer y de 0 à 1. La possibilité d'exécuter les deux ordres simultanément est très improbable, dans la plupart des cas réels on peut s'attendre que soit x, soit y atteigne son seuil de valeur en premier lieu. L'état suivant $\overset{++}{0\ 0}$ n'est donc habituellement pas 11, mais plutôt [10] ou [01]; notons que ce sont des états stables et que dans les deux cas, le 'second' ordre a disparu. On peut représenter ceci comme suit:



Cet exemple nous montre qu'une boucle positive génère de la multistationarité. N'importe quelle boucle positive isolée se comporterait de cette manière. Il n'est pas possible d'exposer ici le détail de ce formalisme, mais cet exemple permet de voir certains de ses principes de base. Une description logique généralisée du système de production de la parole ou de certains de ses composants reste à faire. Cette perspective permet d'entrevoir une meilleure compréhension de la dynamique des systèmes phonologiques.

Remerciements

Les recherches qui sont à la base de ce travail ont été financées par deux subventions du F.N.R.S belge n°: 8.4519.95 dans le cadre de fonds accordés par la loterie nationale et l'autre n°: 1.5.194.97 dans le cadre des 'Crédits chercheurs'. Je tiens particulièrement à remercier Bernard Teston, Alain Soquet et Véronique Lecuit pour leur aide, leur soutien et les discussions que nous avons eu sur divers aspects de ce travail. Thierry Metens et Sergio Hassid m'ont permis d'acquérir les données d'IRM et de pression dans les meilleures conditions. Enfin, je tiens à remercier Jean-Louis Deneubourg, René Thomas, Louis Goldstein, Ian Maddieson, Bjorn Lindblom, René Carré et Jean-Marie Hombert pour les nombreuses et stimulantes discussions que nous avons eu à propos de ces recherches.

Bibliographie

[Bro89] Browman, C. et L. Goldstein (1989), Articulatory gestures as phonological units, Phonology 6: 201-251.

[Bro92] Browman, C. et L. Goldstein (1992), Articulatory phonology: an overview. Phonetica 49: 155-180.

[Car96] Carré, R. (1996), Prédiction des systèmes vocaliques par approche déductive. *Actes des XXIes Journées d'étude sur la parole.* Avignon. pp. 155-158.

[Cho68] Chomsky N. et M. Halle (1968), The Sound Pattern of English, New York, Harper and Row.

[Cla94] Clark, C. et C. Yallop. (1994), Introduction to Phonetics and Phonology, Oxford, Blackwell.

[Con95] Connell, B. et A. Arvaniti, eds., (1995), Papers in Laboratory Phonology 4, Phonetic and Phonological Evidence, Cambridge, Cambridge University Press.

[Dea97] Deacon, T. (1997), The Symbolic Species. The co-evolution of language and the human brain, London, The Penguin Press.

[Dem88] Demolin, D. (1988), Some problems of phonological reconstruction in Central Sudanic. In M. Dominicy and J. Dor (eds.). *Phonological Reconstruction: Problems and Methods. Belgian Journal of Linguistics, 3.* 53-95.

[Dem96a] Demolin, D. et B. Teston (1996), Aspects articulatoires et aérodynamiques des occlusives labio-vélaires. *Actes des XXIèmes Journées d'Etude de la Parole.* Avignon.

[Dem96b] Demolin, D. et B. Teston (1996), Labiodental flaps in Mangbetu, Journal of the International Phonetic Association, 26, 2, pp. 103-111.

[Dem96c] Demolin, D., M. George et A. Soquet (1996), The role of self-organisation in the emergence of phonological systems, Communication à la conference 'The Evolution of Language', Edinburgh.

[Dem97a] Demolin, D. et B. Teston (1997), Phonetic characteristics of double articulations in some Mangbutu Efe languages. *Proceedings of Eurospeech.*, Rhodes.

[Dem97b] Demolin, D., M. George, V. Lecuit, T. Metens, B. Raeymaeckers et A. Soquet (1997), Coarticulation and articulatory compensations

studied by dynamic MRI. *Proceedings of Eurospeech.*, Rhodes.

[Den89] Deneubourg, J.-L. et S. Goss, (1989), Collective patterns and decision making, Ethology, Ecology and Evolution, pp. 295-311.

[Dim86] Dimmendaal, G. J. (1986), Language typology, comparative linguistics and injective consonants in Lendu, Afrika und Überzee 69, pp. 161-192.

[Doc92] Docherty G. J. et D. R. Ladd eds., (1992), Papers in Laboratory Phonology 2, Gesture, Segment, Prosody, Cambridge, Cambridge University Press.

[Don91] Donald M. (1991), *Origins of the Modern Mind. Three stages in the Evolution of Culture and Cognition*. Caambridge Mass. Harvard University Press.

[Dur90] Durand J. (1990), Generative and Non-Linear Phonology, London, Longman.

[Fow80] Fowler, C. et M. Turvey (1980), Immediate compensation in bite-block speech, Phonetica, 37, pp. 306-326.

[Gol95] Goldsmith J. (1995), The Handbook of Phonological Theory, Oxford, Blackwell.

[Gro97] Grosjean, F. et U. Fraunefelder (1997), A guide to the spoken Word Recognition Paradigms, Special Issue of Language and Cognitive Process, Hove, Psychology Press.

[Haw92] Hawkins, S. (1992), An introduction to task dynamics, In Docherty G. J. et D. R. Ladd eds., Papers in Laboratory Phonology 2, Gesture, Segment, Prosody, Cambridge, Cambridge University Press, pp. 9-25.

[Kea94] Keating, P. (1994), Papers in Laboratory Phonology 3, Phonological Structure and Phonetic Form, Cambridge, Cambridge University Press.

[Kel95] Kelso J. A. (1995), *Dynamic Systems. The Self-Organization of Brain and Behavior.* Cambridge. Bradford, MIT Press.

[Kel86] Kelso, J. A., E. L. Saltzaman et B. Tuller (1986), The dynamical perspective on speech production: data and theory. Journal of Phonetics, 14, pp.29-59.

[Ken81] Kent, R. D. (1981), Articulatory-acoustic perspectives on spech development, In R. Stark ed., Language Behaviour in Infancy and Early Childhood, New York, Elsevier, North Holland.

[Ken84] Kent, R. D. (1984), Psychobiology of speech development: Coemergence of language and a movement system, American journal of Physiology, 246, R888-94.

[Kin90] Kingston J. et M. Beckman eds., (1990), Papers in Laboratory Phonology 1, Between the Grammar and the Physics of Speech, Cambridge, Cambridge University Press.

[Lad71] Ladefoged, P. (1971), Preliminaries to linguistic phonetics, Chicago, CUP.

[Lad96] Ladefoged, P. et I. Maddieson (1996), The Sounds of the world's Languages, Oxford, Blackwell.

[Lin79] Lindblom, B., J. Lubker and T. Gay (1979), Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation, Journal of Phonetics, 7, pp. 147-162.

[Lin83] Lindblom, B., P. MacNeilage et M. Studdert-Kennedy (1983), Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie and O. Dahl (eds.). *Explanation for Language Universals.* Berlin. Mouton. 181-203.

[Lin86] Lindblom, B. (1986), Phonetic universals in vowel systems. In J. J. Ohala and J. J. Jaeger (eds.) *Experimental Phonology.* Orlando. Academic Press.

[Lin88] Lindblom, B. and I. Maddieson. (1988), Phonetic universals in consonant systems. In L. Hyman and C. Li (eds.) *Language, Speech and Mind.* London. Routledge.

[Lin92] Lindblom, B. (1992), Phonological units as adaptive of lexical development, In C.S. Ferguson, L. Menn et C. Stoel-Gammon eds., Phonological Development: Models, Research, Implications, Timonium, York Press.

[Lub79] Lubker, J. (1979), The reorganization time of bite-block vowels, Phonetica, 36, pp. 273-293.

[Mad84] Maddieson, I. (1984), Patterns of Sounds, Cambridge, Cambridge University Press.

[Mal66] Malécot, A. (1966), The effectiveness of intraoral air pressure pulse parameters in distinguishing between stop cognates, Phonetica, 14, pp. 65-81.

[Mal70] Malécot, A. (1970), The lenis-fortis opposition: Its physiological parameters, Journal of the Acoustical Society of America, 47, 1588-1592.

[Moo93] Moon, J. B. J. W. Folkins et A. Smith (1993), Air pressure regulation during speech

production, Journal of the Acoustical Society of America, 94, pp. 54-63.

[Mül70] Müller, E. M. et W. S. Brown (1980), Variations in the supraglottal air pressure waveform and their articulatory interpretation, In N. J., Lass ed., Speech and language: advances in basic research and practice, New-York, Academic Press, pp. 317-389.

[Oha76] Ohala J. J. (1976), A model of speech aerodynamics, Report of the phonology laboratory (Berkeley), 1, pp. 93-107.

[Oha86] Ohala, J. J. et J.J. Jaeger (1986), Experimental Phonology, Orlando, Academic Press.

[Oha90] Ohala, J. J. (1990), The phonetics and phonology of aspects of assimilation, in J. Kingston et M. Beckman eds., Papers in laboratory Phonology 1, Cambridge, Cambridge Unievrsity Press, pp. 258-275.

[Oha95] Ohala, J. J. (1995), Experimental phonology, In J. Goldsmith ed. The Handbook of Phonological Theory, Oxford, Blackwell, pp. 713-722.

[Pop59] Popper, K. (1959), The logic of scientific discovery, London, Hutchinson.

[Rem84] Remacle, L. (1984), La différenciation des géminées *mm, nn* en *mb, nd*, Paris, Les Belles Lettres.

[Rou08] Rousselot P. J. (1897-1908), Principes de phonétique expérimentale, Paris, Welter.

[Sal86] Saltzman, E. L. (1986), Task Dynamics coordination of the speech articulators: a preliminary model, In H. Heuer et C. Fromm eds., Generation and Modulation of Action Patterns, New york, Springer, pp. 129-144.

[Sal89] Saltzman, E. L. et K. G. Munhall (1989), A dynamical approach to gestural patterning in speech production, Ecological Psychology, 1: 333-382.

[Sch97a] Schwartz, J.-L., Boë, L.-J., Vallée, N. et C. Abry, (1997a), Major trends in vowel system inventories, Journal of Phonetics, 25, 3, pp. 233-253.

[Sch97b] Schwartz, J.-L., Boë, L.-J., Vallée, N. et C. Abry, (1997b), The dispersion focalization theory of vowel systems, Journal of Phonetics, 25, 3, pp. 255-285.

[Sno93] Snoussi, E. H. et R. Thomas (1993), Logical identification of all steady states: the concept of feedback loop characteristic states, *Bulletin of Mathematical Biology 55*, pp. 973-991.

[Sto96] Stone, M. (1996), Instrumentation for the study of speech physiology, , In N. J. Lass, Principles of Experimental Phonetics, St Louis, Mosby, pp. 495-524.

[Sto97] Stone, M. (1997), Laboratory techniques for investigating sppech articulation, In W. J. Hardcastle et J. Laver eds., The Handbook of Phonetic Sciences, Oxford, Blackwell, p. 11-32.

[Stu87] Studdert-Kennedy. M. (1987), The Phoneme as a Perceptuomotor Structure. In A. Allport, D. MacKay, W. Prinz and E. Sheerer (eds.) *Language perception and Production*. London. Academic Press. 67-84.

[Tes90] Teston, B. et B. Galindo (1990), Design and development of a work station for speech production and analysis. Proc. VERBA 90, International Conference on Speech Technologies, Rome, pp. 400-408.

[The89] Thelen, E. (1989), Self-organisation in developmental processes: Can systems approaches work? In M. R Gunnar et E. Thelen eds., Systems and Development, Hillsdale, New Jersey, Lawrence Earlbaum.

[The91] Thelen, E. (1991), Motor aspects of emergent speech; a dynamic approach. In N. A. Krasnegor, D. M. Rumbaugh, R. L. Schiefelbusch et M. Studdert-Kennedy eds., Biological and Behavioural Determinants of Language Development, Hillsdale, New Jersey, Lawrence Earlbaum.

[Thé97] Théraulaz, G., et F. Spitz, eds., (1997), Auto-organisation et comportement, Paris, Hermes

[Tho90] Thomas, R. et R. D'Ari (1990), Biological Feedback, Boca Raton, Florida, CRC Press.

[Tho91] Thomas R. (1991), Regulatory Networks Seen as Asynchronous Automata: A Logical Description. *Journal of Theoretical Biology 153*. 1-23.

[Tho95] Thomas R., D. Thieffry and M. Kaufman. (1995), Dynamical behaviour of biological regulatory networks-I. Bilogical role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of Mathematical Biology 57, 2*, pp. 247-276.

[Tra95] Traill, A. (1995), The perception of clicks in !xoo , Journal of African Languages and Linguistics, 15, pp. 161-174.

[Tra97] Traill, A. et R. Vossen (1997), Sound change in Khoisan languages: new data on click loss and replacement, Journal of African Languages and Linguistics, 18, pp. 21-56.

[Tuc56] Tucker, A. N. et M. A. Bryan (1956), The non-Bantu Languages of North-Eastern Africa, London, OUP.

[Val94] Vallée, N. (1994), *Systèmes vocaliques: de la typologie à la prédiction.* Doctoral dissertation. Université de Grenoble.

[Vör74] Vörbichler, A. (1974), Die Sprache der Mamvu, Glückstad, Augustin.

[War80] Warren, D. W., G. Nelson et G. D. Allen (1980), Effects of increased vertical dimension on oral port size and fricative intelligibility, Journal of the Acoustical Society of America, 67, pp. 1828-1831.

[War81] Warren, D. W., D. J., Hall et J. Davis (1981), Oral port constriction and pressure airflow relationship during sibiliants productions, Folia Phoniatrica, 33, pp. 380-394.

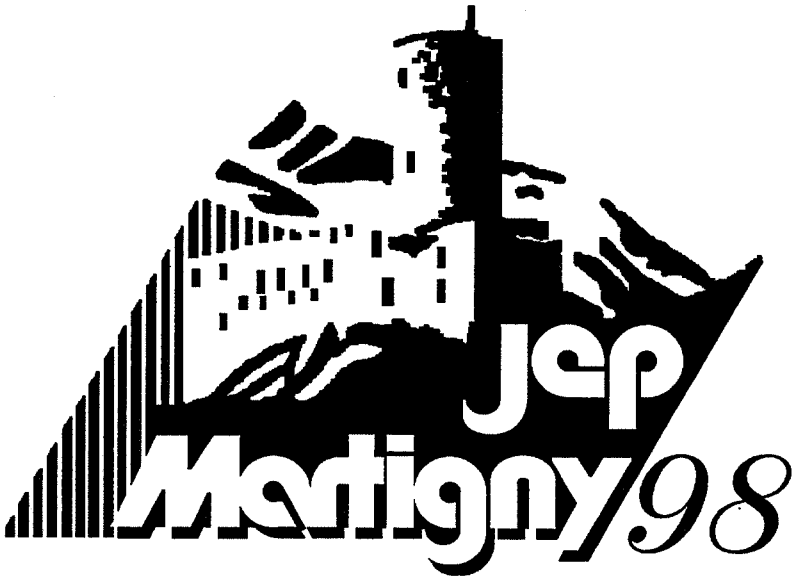
[War82] Warren, D. W. (1982), Aerodynamics of Speech production, In N. J. Lass L. V. McReynolds J. L. Northern et D. E. Yoder eds. Speech Language and Hearing vol. 1., Philadelphia, Saunders, pp. 219-245.

[War84] Warren, D. W., G. Allen et H. A. King (1984), Physiologic and perceptual effects of induced anterior open bite, Folia Phoniatrica, 36, pp. 164-173.

[War96] Warren, D. W. (1996), Regulation of speech aerodynamics, In N. J. Lass, Principles of Experimental Phonetics, St Louis, Mosby, pp. 46-92.

[Wyk81] Wyke, B. (1981), Neuromuscular control systems in voice production. In D. Bless ed., Vocal fold physiology: contemporary research and clinical issues, San Diego, College-Hill Press, pp. 71-76.

Phonétique I



Identification dialectale des parlers arabes et détermination expérimentale d'indices discriminants.

Melissa Barkat

Laboratoire Dynamique Du Langage
M.R.A.S.H
14, avenue Berthelot
69363 Lyon Cedex 07
Tel : 04 72 72 64 63 / Fax : 04 72 72 65 90
Serveur : http://www.ddl.mrash.fr
e-mail : Melissa.Barkat@mrash.fr

Abstract

This paper deals with the identification of spoken Arabic based on natural speech. Six Arabic dialects were chosen from the major geographical subdivision of the Arab World (i.e. Maghreb vs. Middle-East). Speech data were obtained by recording spontaneously produced narration (i.e.description of pictures) of 12 Arabic speakers. Samples of speech (complete utterances) were then selected so as to make up the stimulus material for a perception test. The results of the experiment show that Arabic listeners are successful in identifying the dialect used by Arabic speakers thanks to several cues of different nature (segmental, syntactic, lexical and prosodic). This experiment must be regarded as a first step towards the detection of reliable cues for the Automatic Identification of Arabic Dialects.

1. Introduction

La plupart des systèmes de reconnaissance linguistique s'appuient sur les formes standardisées des langues. Or, un grand nombre de langues naturelles se déclinent sous des formes dialectales plus ou moins "ressemblantes".

Certains parlers arabes, par exemple, présentent si peu de caractéristiques linguistiques communes avec la langue standard que les rapprochements sont parfois difficiles à établir. Il convient donc, dans le cadre des recherches sur l'Identification Automatique des Langues[1], d'élargir le champ des investigations aux parlers non-standards afin d'optimiser les résultats dans ce domaine. La présente étude constitue la première étape vers la détermination "d'indices acoustiques robustes" [Hom1997] pour l'identification automatique des parlers arabes. Elle se base sur les résultats d'une expérience perceptuelle à partir de parole naturelle menée auprès d'arabophones originaires de différents pays du Maghreb et du Moyen-Orient.

2. Géographie dialectale du Monde Arabe

Chaque parler arabe présente des caractéristiques qui lui sont propres. Celles-ci peuvent être d'ordre syntaxique, lexical, segmental et prosodique. Certains parlers ont "innové" à l'intérieur de chacun de ces domaines et ne présentent que peu de caractéristiques communes avec l'arabe classique (i.e.standard).

Une vue d'ensemble du domaine linguistique arabe, permet de distinguer deux zones à l'intérieur du continuum dialectal : à l'ouest, la zone occidentale regroupant les parlers du Maghreb ; à l'est, la zone orientale constituée des parlers Moyen-Orientaux [Coh68] [Fle75]. A cette première distinction purement géographique, il convient de superposer une subdivision secondaire qui distingue les parlers citadins des parlers ruraux [Coh68]. Les parlers arabes constituent ainsi des formes linguistiques identifiables pour "l'oreille" arabophone.

3. Corpus et Méthode

Six variétés dialectales représentatives des deux grandes zones géographiques du Monde Arabe ont été sélectionnées (table 1). Une base de données acoustiques à été élaborée par l'enregistrement de 12 locuteurs (et locutrices) arabophones s'exprimant de manière spontanée dans leur dialecte maternel à partir d'un livre d'images sans texte[2].

Table 1. Origines dialectales représentées à l'intérieur des stimuli et chez les sujets testés.

Maghreb	Moyen-Orient
Maroc	Syrie
Algérie	Liban
Tunisie	Jordanie

Les enregistrements, effectués en chambre insonorisée sur magnétophone DAT, ont été digitalisés à 22 kHz,16 bits, mono. Quatre-vingt-seize échantillons de parole (énoncés complets) de durée variable (entre 5 et 30 secondes) ont ensuite été extraits du corpus et présentés, en ordre aléatoire, comme stimuli lors d'un test d'identification dialectale mené auprès de 18 autres sujets arabophones originaires des six mêmes pays et ayant pour tâches les trois points suivants :

a) d'identifier l'origine dialectale du locuteur entendu en fonction de la zone (i.e. Maghreb vs Moyen-Orient).

b) d'identifier le pays dont pouvait être originaire le locuteur parmi les 6 choix proposés (table 1).

[1] Ce travail entre dans le cadre d'un projet sur la discrimination multilingue mené en collaboration avec l'ILPGA (Paris III), l'IRIT (Toulouse) et l'ICP (Grenoble) et financé par la Direction de la Recherche et de la Technologie (DRET-DGA). Contrat n° 95/118.

[2] Mayer, M., 1980, Frog, where are you ?, Dial Books for Young Readers, New-York :

c) de définir (dans la mesure du possible) les indices prosodiques, segmentaux et/ou lexicaux ayant permis l'identification.

Outre la détermination d'indices discriminants à priori pertinents pour l'identification des parlers arabes, nous avons voulu vérifier auprès de nos sujets les hypothèses suivantes :

1. L'identification par zone est aisée pour l'ensemble des sujets.

2. Les meilleurs scores d'identification sont obtenus pour la distinction dialecte maternel / autre dialecte.

3. Les résultats les moins probants concernent les parlers proches à l'intérieur d'une même zone.

4. Résultats et Discussion

4.1. Identification dialectale

4.1.1. Identification par zone

Les résultats de l'identification dialectale par zone révèlent que 97% des stimuli maghrébins et 99% des stimuli moyen-orientaux ont été identifiés correctement.

Ces chiffres confirment l'idée que la partition dialectale du domaine linguistique arabe par zone (Maghreb vs Moyen-Orient) transparaît à travers des éléments perceptuels identifiables par l'ensemble des locuteurs arabophones toutes origines confondues.

4.1.2. Identification par pays et première approche d'analyse des erreurs

Les figures 2 illustrent le taux d'identification par pays obtenu par les 18 sujets.

Nos résultats, bien qu'asymétriques d'une zone à l'autre, montrent que les meilleurs scores sont obtenus :
a) pour l'identification du dialecte maternel par rapport aux autres parlers.
b) pour l'identification des variétés dialectales appartenant à la même zone géographique que le parler d'origine.

Toutefois, il convient de remarquer que les sujets maghrébins présentent un taux d'identification par pays homogène (i.e. bonne localisation des parlers maghrébins et moyen-orientaux) par rapport aux sujets moyen-orientaux, ces derniers distinguent, en effet, assez mal les parlers maghrébins entre eux.
Les mauvaises identifications concernent majoritairement des parlers proches à l'intérieur d'une même aire géographique (Maroc et Algérie au Maghreb ; Syrie et Liban au Moyen-Orient). Elles correspondent à des zones où les frontières dialectales et nationales ne coïncident pas. [Len94] et [Mar77].

Ces parlers, attestent en effet, de nombreux traits communs (comme, entre autres, la réalisation glottale de [ɋ]) en Syrie et au Liban ; ou la simplification du système vocalique des parlers maghrébins), ce qui expliquent les difficultés rencontrées par les sujets pour la tâche (b). Une analyse plus fine des erreurs et des hésitations est en cours et sera présentée ultérieurement.

Fig.2(a) Stimuli Marocains

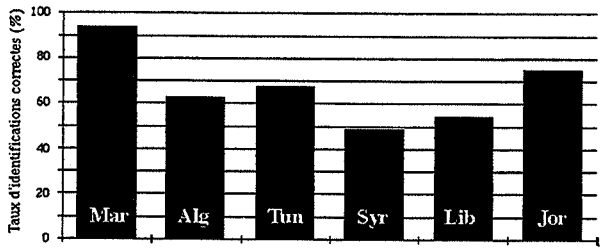


Fig.2(b) Stimuli Algériens

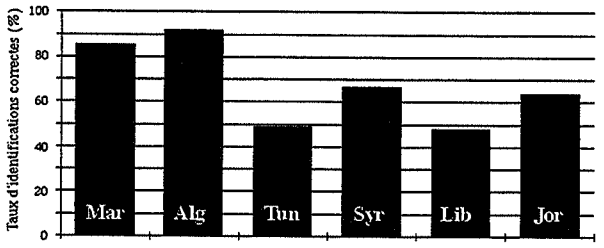


Fig.2(c) Stimuli Tunisiens

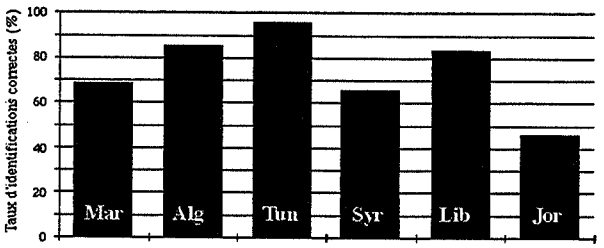


Fig.2(d) Stimuli Syriens

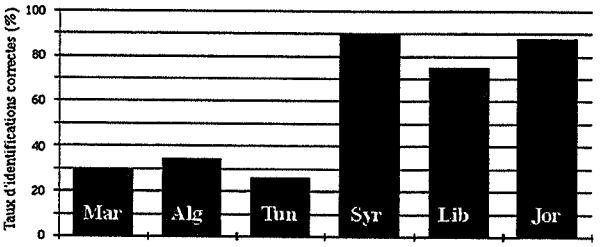


Fig.2(e) Stimuli Libanais

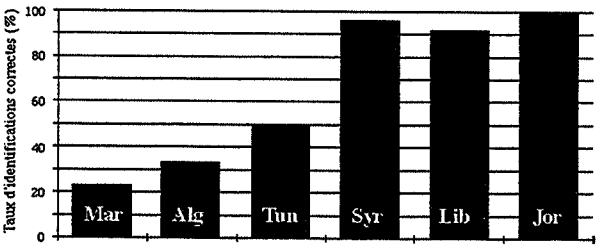


Fig.2(f) Stimuli Jordaniens

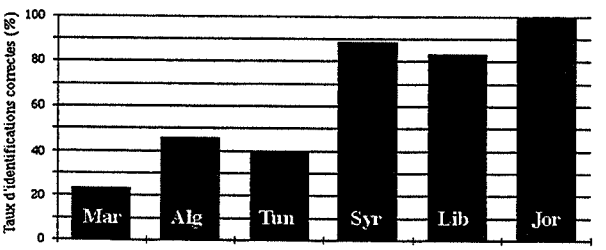


Fig.2 Identification dialectale par pays en fonction de l'origine des sujets (résultats)

4.2. Détermination expérimentale d'indices discriminants pour les parlers arabes

Le dialecte maternel des sujets ayant participé à l'expérience constitue le filtre au travers duquel ils ont repéré les particularités des autres parlers. Les indices cités ci-dessous peuvent être considérés comme pertinents pour la reconnaissance automatique des parlers arabes dans la mesure où ils sont perçus par les locuteurs/auditeurs arabophones comme les éléments permettant l'identification.

A partir de tous les critères discriminants évoqués, nous avons établi une classification par type d'indices (phonétiques, morpho-syntaxiques et lexicaux). Il convient cependant d'attribuer à ces éléments une valeur relative (i.e. localement limitée aux parlers de nos informateurs[3]).

Pour avoir une idée de leur représentativité et notamment de l'extension géographique de ces traits en dehors de nos points d'enquête, on pourra se reporter, entre autres, aux références citées.

4.2.1. Indices phonétiques

Réalisation de la plosive uvulaire sourde [q]
(sonore vs sourde vs glottalisation)
cf. [Can60] [Coh70] [Gha77] [Fer69].
[g] (Jordanie) ; [q] (Maroc, Algérie et Tunisie) ; [/] (Syrie et Liban).
[g r o n] vs [q r o n] vs [/ o r n] "cornes"

Affrication de la plosive dentale [t].
[t]>[t s] en Algérie (Cherchell).
[t k « s « r]>[t s k « s « r] "il s'est cassé"

Traitement des interdentales [T ; D ; D‡] (conservation ou confusion avec les plosives ou les fricatives labio-dentales correspondantes) [TaiPrep].
[T] en Algérie et en Jordanie ; [t] au Maroc ; [f] à Bousalem (Tunisie).
[D‡] en Jordanie ; [d‡] au Liban.
[T « m a] vs [t « m a] vs [f « m a] "alors"
[D ‡A f D ‡A ? A] vs [d ‡A f d ‡A ? A]"grenouille"

Affrication ou vélarisation de la fricative post-alvéolaire [Z] [Kay79].
[Z]>[g] à Rabat (Maroc).
[Z]>[t S] à Alep (Syrie).
[Z]>[d Z] en Algérie
[Z ˇ l « s]>[g ˇ l « s] "assis"
[w a Z h o]>[w a t S o] "son visage"
[Z r a n a] [d Z r a n a] "grenouille"

Sonorisation contextuelle (régressive) de la consonne emphatique [s‡]>[z‡] pour l'ensemble de nos parlers orientaux.
[s‡ i r]>[z‡ i r] "petit"

Sonorisation de la plosive vélaire sourde [k]>[g] en Syrie (Alep).
[b j « r k o d ‡]>[b j « r g o d ‡] "il court"

Passage d'une liquide à une nasale [l] >[n] au Liban (Beyrouth).
[m l i ' e]>[m n i ' e] "belle"

Altération conditionnée de la fricative post-alvéolaire sourde [S]>[s] devant [Z] en Algérie [Tai84].
[S « Z r a]>[s « Z r a] "arbre"

Divergences des voyelles thématiques (i.e. correspondant à une forme syntaxique).
[h a r a b]>[h i r i b] 'il s'est sauvé" au Liban (Zahlé)

Centralisation ou chute (notée [•]) des voyelles brèves dans les parlers du Maghreb [Bar97].
[s a k a t]>[s• k « t] "il s'est tu"
[h a r a b ˇ t]>[h a r• b « t] "elle s'est sauvée"

Antériorisation de la voyelle ouverte (i.e. "imala") en position finale et médiane [a]>[e] en Tunisie (dans les monosyllabiques uniquement) et pour l'ensemble de nos parlers orientaux [Bar97].
[l m e] "eau" ;
[' e l o] "son état" ;
[n d ‡i f e] "propre"

Passage d'une diphtongue à une voyelle longue [a j]>[e] au Liban [i] en Syrie.
[m ˇ f i S i] "il n'y a rien" ;
[b e l l e S e] "sans rien"

Diphtonguaison d'une voyelle longue [i]>[e j] et ; [a]>[a j j] au Moyen-Orient.
[b ˇ ? d i n]>[b ˇ ? d e j n] "ensuite"
[ˇ l m a /]>[ˇ l m a j j] "l'eau".

4.2.3. Indices morpho-syntaxiques

Formes de l'inaccompli (présence vs absence d'une particule préfixale en début d'item verbal).
Absence de particule en Algérie et en Tunisie, présence du préfixe [k -] au Maroc et de [b -] au Moyen-Orient.
[i ? i j « t ‡] "il crie"
[k i « w w t] "il aboit"
[b j « r g o d ‡] "il court"

Chute du pronom personnel affixe de 3°pers. [- **h**o] "son" [- **h**a] "sa" et allongement compensatoire de la voyelle finale en Syrie, au Liban et en Jordanie.
[/ o r n o] "ses cornes (à lui)"

Infixation de la séquence [- i j j «-] pour la forme diminutive en Algérie (au lieu de [-ej-] partout ailleurs).

[k « l b] "chien"	[k l i j j « b] "petit chien"
[w l « d] "garçon"	[w l i j j « d] "petit garçon

4.2.4. Indices lexicaux

Expression du rapport d'annexion :
[n t ˇ ? o] "le sien ; son" à Jijel et [n t s ˇ ? o] à Cherchell (Algérie).
[b t ˇ ? o] "le sien ; son" en Jordanie et en Syrie.
[d i ˇ l o] "le sien ; son" au Maroc.
[t ‡A b a ? o] "le sien, son" au Liban

Expression de l'indéfini :
[h o n k a l a b] "un chien"
[w a ' « d l k « l b] "un chien"

[3]Rabat et Safi (Maroc), Jijel et Cherchell (Algérie), Bizerte et Bousalem (Tunisie), Homs et Alep (Syrie), Beyrouth et Zahlé (Liban), Irbid (Jordanie).

4.2.5. Indices suprasegmentaux et prosodiques

Emphase appliquée à un seul segment (Tunisie) ou à tout le mot (Jordanie).

Rythmes et processus accentuels différents entre parlers maghrébins et moyen-orientaux (i.e. organisation syllabique différente).

Trait de nasalité constante au Liban (la vérification de cet indice est en cours d'analyse acoustique).

5. Conclusion

Au terme de cette expérience d'identification inter-dialectale et de distinction à l'intérieur des zones linguistiques ainsi définies, nous sommes en mesure de tirer les conclusions suivantes :

Le domaine linguistique du monde arabe est caractérisé d'une part, par une certaine hétérogénéité (i.e. opposition parlers occidentaux vs. parlers orientaux), plusieurs critères acoustiques perceptuellement significatifs permettent d'établir cette distinction.

A l'intérieur de ces deux grandes zones, caractérisées par des indices comparables, une distinction plus fine est souvent possible, chaque parler étant constitué de quelques traits exclusifs qui, s'ils sont perçus, permettent une identification locale précise.

En revanche, certaines zones dialectales présentent des frontières floues au regard des délimitations géographiques en terme de "pays". Elles autorisent un regroupement dialectologique interne qui a donné lieu à de nombreuses études. Cependant, dans le cadre de la reconnaissance automatique il est nécessaire de se concentrer sur la détermination de points divergents. Quelques critères d'identification ont été expérimentalement déterminés dans cette étude préliminaire, certains d'entre eux feront l'objet d'analyses plus poussées et accéderont peut-être au rang d'indices acoustiques robustes pour l'identification automatique des parlers arabes.

6. Références

[Bar97] Barkat M. & al. (1997), «Détermination d'indices acoustiques robustes pour l'identification des parlers arabes : Etats d'[a]mes vocaliques», in Actes des Journées d'Études Linguistiques, (5-6 décembre 1997), pp.141-146, Nantes :

[Can60] Cantineau J (1960), Cours de phonétique arabe, Librairie C. Klincksieck, Paris :

[Coh68] Cohen D. (1968), "Langue Arabe" in Encyclopédie Universalis (T.II) pp. 201, Paris :

[Coh70] Cohen D. (1970), "Koiné, langue commune et dialectes arabes", Janua Linguarum, Série Practica vol 81, pp.105-25, Mouton, The Hague :

[Fer69] Ferguson C.A (1969), "The /g/ in Syrian Arabic : filling a gap in a phonological pattern", in Word, vol.25, pp.114-19.

[Fle75] Fleish H. (1975), "Arabiyya" in Encyclopédie de l'Islam (T.I) pp. 593, Brill & Maisonneuve Ed., Leiden, Paris :

[Gha77] Ghazali S., (1977), "On the controversy of the Arabic <qaf>", in Texas Linguistic Forum, vol.6, pp.40-51, Austin, Texas :

[Hom97] Hombert J.M & I. Maddieson (1997), «A linguistic approach to Automatic Language Recognition», in Actes du Congrès International des Linguistes, (Juillet 1997), Paris : (sous presse).

[Kay79] Kay A.S (1979), "Arabic <jiim> a synchronic and diachronic study" in Linguistics, pp.31-72.

[Len94] Lentin J. (1994), "Classification et typologie des dialectes du Bilâd-al-Sam. Quelques suggestions pour un réexamen", MAS-GELLAS nouvelle série, n°6, pp.11-42, Paris :

[Mar77] Marçais Ph. (1977), Esquisse grammaticale de l'arabe maghrébin, Maisonneuve, Paris :

[Tai84] Taine-Cheikh C. (1984), "Les altérations conditionnées des chuintantes et des sifflantes dans les dialectes arabes", in Compte-rendu du Groupe de Linguistique Chamito-Sémitiques (T.XXIV-XXVIII), pp. 413-435, Geuthner, Paris :

[TaiPrep] Taine-Cheikh, C. (en préparation) "Les interdentales dans les dialectes arabes : un macro discriminant".

Classement automatique de phonèmes dans un cadre multilingue

C. Corredor-Ardoy, P. Boula de Mareüil, M. Adda-Decker, L. Lamel, J.L. Gauvain

LIMSI-CNRS, BP 133
91403 Orsay Cedex, FRANCE
Tél.: ++33 1 69 85 81 02 - Fax: ++33 1 69 85 80 88
e-mail: {corredor, mareuil, madda, lamel, gauvain}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this article, we describe an approach for automatic multi-lingual phoneme classification. The classes were obtained by agglomerative hierarchical clustering. We used a similarity measure based on the likelihood between the acoustic frames and the Hidden Markov Models. The method was applied to French, English, German, Spanish (IDEAL corpus), as well as to Italian and Portuguese (SPEECHDAT corpus). The analysis of the clusters demonstrated that, despite the acoustic mismatch between these corpora, this approach remains robust. For 90 clusters, the obtained classes correspond, to a large extent, to well defined linguistic groups. A qualitative analysis of the results is given.

1. INTRODUCTION

La recherche d'une typologie des phonèmes constitue un des enjeux de l'étude de la parole [9, 17] dans un but d'enseignement [5] ou de traitement automatique de la langue [2, 3]. Dans cet article, nous décrivons un algorithme de classement visant à regrouper les phonèmes de six langues européennes (français, anglais, allemand, espagnol, italien et portugais), et nous donnons une interprétation linguistique des résultats.

Le classement automatique des phonèmes dans un cadre multilingue a été incité par le développement de corpus destinés à l'identification automatique de la langue (IAL) : OGI-TS [13], CALLHOME, CALLFRIEND, SPEECHDAT, IDEAL [4]... Il est possible de remplacer plusieurs systèmes de reconnaissance phonétique par un seul système commun et unique à toutes les langues [4].

Dans notre approche, l'analyse acoustique est fondée sur les coefficients cepstraux plutôt que sur l'extraction de formants. Cette dernière en effet nécessite une segmentation voisé/non voisé, la localisation des noyaux vocaliques et l'extraction de la valeur des formants, décisions qui sont sujettes à erreur.

Nous proposons un algorithme de classement hiérarchique. Berkling [2] a également appliqué cette méthode à l'IAL de 6 langues du corpus OGI, mais elle utilise une mesure de similitude fondée sur des distances entre vecteurs acoustiques. Köhler [11] utilise lui une mesure définie sur des vraisemblances acoustiques ; néanmoins, cette mesure a seulement été apliquée à 3 des 11 langues du corpus OGI. La normalisation introduite dans notre mesure de similitude, de même fondée sur des vraisemblances acoustiques, a permis d'appliquer le classement automatique dans un cadre plus étendu : celui des corpus IDEAL [4] et SPEECHDAT. La mesure utilisée s'est révélée robuste par rapport aux différentes conditions d'enregistrement et au contenu liguistique des corpus.

Dans un cadre monolingue, l'algorithme de classement hiérarchique peut aider à l'analyse des classes phonétiques, à la définition de l'ensemble des unités destiné à la reconnaissance de la parole, et au groupement d'un grand nombre de modèles markoviens en contexte.

La section 2 présente l'algorithme de classement automatique. La section 3 est consacrée au cadre expérimental : corpus et alphabets. La section 4 analyse les résultats obtenus en faisant diminuer le nombre de classes à partir des 235 phonèmes de départ, et détaille le regroupement en 90 classes, avant de conclure.

2. L'ALGORITHME DE CLASSEMENT AUTOMATIQUE

2.1. *Préliminaire*

Depuis l'apparition des premiers systèmes de reconnaissance vocale fondés sur une modélisation phonétique markovienne, il s'est avéré important de pouvoir mesurer la distance entre les modèles. Cette mesure est utilisée pour réduire le nombre des modèles en contexte, pour déterminer l'ensemble optimal des phonèmes dans une langue ou pour établir un ensemble phonétique commun à plusieurs langues. Parmi les algorithmes de classement automatique, l'algorithme *k*-moyennes [6] est très dépendant de l'initialisation des classes. Nous avons ici opté pour une arborescence hiérarchique, même si celle-ci n'autorise pas de retour en arrière (même mal calculée, une classe ne peut être reconsidérée).

Les phonèmes des six langues ont ainsi été groupés par classement automatique hiérarchique [6]. Dans la phase d'initialisation de l'algorithme, chaque phonème est assigné à une classe. Après chaque itération, les deux classes de mesure de similitude maximale sont regroupées. La procédure se répète jusqu'à l'obtention du nombre de classes désiré. À noter que certaines classes peuvent contenir plusieurs phonèmes d'une même langue, car aucune contrainte sur l'origine linguistique des phonèmes n'a été prise en compte[1]

[1] Notre approche entre donc dans le type de classement maximal (inter-langue + intra-langue) proposé par Berkling [3].

2.2. *La mesure de similitude*

La définition d'une mesure de similitude (ou dissimilitude) entre phonèmes ou allophones est un sujet qui a été largement traité par la communauté scientifique. Young [18] a défini la dissimilitude entre allophones en exprimant la divergence de deux gaussiennes en fonction de leur moyenne et de leur variance. Cependant, cette approche est seulement applicable à des modèles markoviens à un seul état et une seule gaussienne. En remplaçant le concept de divergence par la distance de Bhattacharyya, Mak [12] a proposé une expression alternative de la dissimilitude en fonction des paramètres des modèles. Cette approche est applicable à des modèles markoviens avec mélange de gaussiennes, mais à un seul état.
La comparaison de deux modèles markoviens à plusieurs états et mélange de gaussiennes requiert une mesure de similitude (ou dissimilitude) fondée sur la vraisemblance des données acoustiques (coefficients cepstraux) par rapport aux modèles markoviens. Dans cette ligne de travail, Juang [10] et Köhler [11] ont proposé une mesure de dissimilitude définie comme la différence des vraisemblances acoustiques. La mesure ainsi définie est dérivée du concept de divergence entre modèles markoviens.
Dans le cadre du travail présenté dans cet article, nous avons utilisé une mesure de similitude fondée sur le concept d'information mutuelle [14]. Cette mesure constitue une approximation de la probabilité a posteriori $\Pr(\lambda_i|\vec{\varphi}_j)$:

$$S(\varphi_i,\varphi_j) = f(\vec{\varphi}_j|\lambda_i)^\gamma / \sum_{k=1}^{n} f(\vec{\varphi}_j|\lambda_k)^\gamma \qquad (1)$$

où $\vec{\varphi}_j$ correspond aux données acoustiques du phonème φ_j ; λ_i est le modèle markovien du phonème φ_i ; λ_k est le modèle markovien de chacun des phonèmes à classifier ; f est la fonction de densité de probabilité des observations, connaissant les données d'apprentissage ; et n est le nombre d'unités phonétiques. Le coefficient γ a été introduit pour compenser l'hypothèse d'indépendance entre modèles (il a été fixé à 0,5 de façon empirique). Cette mesure étant asymétrique, nous avons utilisé une version symétrique, en prenant la moyenne des mesures de similitude :

$$S_s(\varphi_i,\varphi_j) = \frac{S(\varphi_i,\varphi_j) + S(\varphi_j,\varphi_i)}{2} \qquad (2)$$

Dans le cas où les classes contiennent plus d'un phonème, le concept de mesure de similitude phonétique s'étend à celui de mesure de similitude interclasse. Cette mesure est définie comme la moyenne des similitudes entre phonèmes[2] :

$$S(C_i,C_j) = \frac{1}{n_i n_j} \sum_{\varphi \in C_i} \sum_{\varphi' \in C_j} S_s(\varphi,\varphi') \qquad (3)$$

où n_i et n_j sont les nombres de phonèmes dans les classes C_i et C_j respectivement, et $S_s(\varphi,\varphi')$ est la mesure de similitude phonétique interclasse.

[2]La mesure de similitude entre deux classes peut aussi être calculée comme le maximum ou le minimum des similitudes entre phonèmes [6].

3. CADRE EXPÉRIMENTAL

3.1. *Description des corpus*

L'algorithme de classement hiérarchique a été appliqué aux enregistrements du corpus IDEAL [4] et du corpus SPEECHDAT. Le corpus IDEAL a été conçu pour le développement de systèmes d'identification automatique de la langue. Il contient plus de 300 appels téléphoniques en français, anglais, allemand et espagnol. Le contenu linguistique d'un appel est très varié : 18 phrases lues (phrases de journal, phrases phonétiquement équilibrées, dates, heures, adresses, etc.), 12 questions d'ordre général (code de l'appel, sexe, âge, code postal, etc.) et 6 questions destinées à recueillir de la parole spontanée. Dans le cadre de ce travail, nous avons sélectionné environ 7500 phrases par langue, dont 40% sont lues et 60% constituent des réponses aux questions d'ordre général. Le corpus téléphonique SPEECHDAT a lui été construit pour le développement de systèmes de reconnaissance de la parole dans plusieurs langues européennes. En ce qui concerne l'italien et le portugais, nous avons sélectionné environ 9500 phrases phonétiquement équilibrées par langue.

3.2. *Définition des alphabets phonétiques*

Les unités phonétiques définissant chaque langue ont été sélectionnées différemment. Pour le français, l'anglais et l'allemand, nous avons repris les jeux de phonèmes utilisés par le LIMSI dans le cadre de la reconnaissance de la parole continue de ces langues (34, 44 et 46 phonèmes pour le français, l'anglais et l'allemand respectivement) [7, 1]. Les 24 phonèmes de l'espagnol constituent l'ensemble de base traditionnellement retenu [16]. Enfin pour l'italien et le portugais les jeux phonétiques sont issus de SAMPA [8] (49 et 38 phonèmes pour l'italien et le portugais respectivement). Au total, le nombre d'unités phonétiques est de 235 ; le tableau 2 illustre le nombre de phonèmes par langue et par catégorie.

Table 1: nombre de phonèmes par langue (français, anglais, allemand, espagnol, italien, portugais) et par catégorie (V=voyelles, C=consonnes, syl.=syllabiques).

	Fr.	*An.*	*Al.*	*Es.*	*It.*	*Po.*
V simples	14	14	19	5	7	14
diphtongues	-	6	3	-	-	2
glides ou C syl.	3	2	4	-	2	2
C simples	17	20	20	17	17	20
affriquées	-	2	-	1	4	-
géminées	-	-	-	1	19	-
total	34	44	46	24	49	38

3.3. *Modélisation phonétique*

Chaque phonème est caractérisé au moyen d'un Modèle de Markov Caché à trois états indépendant du contexte (CI-CDHMM: Context Independent Continuous Density Hidden Markov Model). À chaque état, la loi d'émission d'observation est définie comme la somme pondérée de 32 gaussiennes. Pour chacune des langues, l'ensemble des

phrases a été phonétiquement aligné. Cette segmentation a été obtenue en utilisant la transcription orthographique des phrases, le dictionnaire des mots sous forme graphémique et phonémique, et les modèles phonétiques.
À partir des vecteurs cepstraux correspondant aux 235 phonèmes, et à partir des leurs modèles markoviens associés, nous avons calculé la matrice de vraisemblances acoustiques. Avant d'effectuer cette opération, nous avons dû limiter le nombre de vecteurs acoustiques par phonème à moins de 20 000 échantillons (7 000 vecteurs cepstraux en moyenne[3]), pour réduire le temps de calcul de la matrice de vraisemblance. Une fois cette matrice calculée, nous avons appliqué l'algorithme de classement hiérarchique en faisant varier le nombre de classes désirées.

4. ANALYSE DES RÉSULTATS

4.1. *Réduction du nombre de classes*

On peut regrouper de diverses façons les phonèmes d'une langue, a fortiori de plusieurs. Il est classique, en phonétique, d'adopter une représentation sous la forme d'une arborescence plus ou moins hiérarchisée, reflétant un degré d'analyse plus ou moins profond. Ceci est possible avec notre algorithme, où l'on choisit à l'avance le nombre de classes, en faisant décroître ce nombre. Avec 2 classes, on retrouve en gros une bipartition voyelle-consonne ; avec une dizaine, on retombe à peu près sur les groupes suivants, qui s'excluent mutuellement : voyelles ouvertes, voyelles fermées, fricatives, occlusives, nasales, liquides, glides. Même si certains phonèmes sont réfractaires au regroupement, et même si d'autres regroupements ne sont pas ceux que nous souhaiterions, il est intéressant d'observer la pertinence linguistique des classes produites par un calcul sur des traits purement acoustiques (les coefficients cepstraux).
Toutefois, notre but ici se situe à un niveau d'abstraction moindre : nous souhaitons perdre le moins possible d'information phonématique. Nous ne prétendons pas définir formellement ce concept, qu'il est difficile d'évaluer de manière objective. Le phonème et ses allophones sont traditionnellement définis pour une langue donnée, par l'analyse en paires minimales de commutation, liée à la fonction distinctive. Des recommandations et des conventions sont publiées chaque année par des spécialistes, pour décider d'inclure officiellement ou non des symboles et des diacritiques dans l'Alphabet Phonétique International (API). En même temps, ce n'est pas un hasard, intuitivement, si des sons voisins, dans plusieurs langues, sont représentés avec le même symbole API [15]. C'est cette proximité que l'algorithme décrit ci-dessus permet d'explorer objectivement.
Partant des 235 phonèmes de notre ensemble initial (pour le français, l'anglais, l'allemand, l'espagnol, l'italien et le portugais), on a très rapidement des géminées de l'italien qui se regroupent avec les consonnes simples correspondantes, mais ce n'est que vers 90 classes qu'émerge une classe /t/, commune aux 6 langues. Avec 80 classes, les /z/ se retrouvent regroupés, mais le /o/ est rattaché au /u/. Avec la part d'arbitraire que cela comporte (mais il faut trouver un compromis), c'est le regroupement en 90 classes que nous allons maintenant décrire plus en détails. De surcroît, 90 est grosso modo le nombre de symboles API différents qui sont utilisés pour décrire nos 6 langues.

4.2. *Interprétation des 90 classes de phonèmes*

On compte 35 singletons (correspondant à presque autant de voyelles que de consonnes) : des exemples en sont le /ɥ/ français, le /ð/ et le /ʌ/ anglais, le /ç/ allemand, qui sont propres à leurs langues. Parmi les 55 classes restantes, reportées dans le tableau 2, certaines rassemblent uniquement 6 phonèmes de symbole API identique à travers les 6 langues (plus un 7^{e} pour les consonnes géminées de l'italien) : /i/, /s/, /m/, /t/, /p/, /k/. Dans ces chiffres, il faut bien sûr compter avec le bruit introduit initialement par des jeux de phonèmes hétérogènes (notamment en ce qui concerne les traits de durée, les diphtongues, les glides, les affriquées, les géminées), plus ou moins fins en allophones (vocaliques). Les affriquées et les diphtongues se voient souvent isolées - même si les /aɪ/ allemand et anglais sont regroupés. Notre modèle ne permet pas de déterminer s'ils s'apparieraient avec 2 segments (par exemple /t/ et /ʃ/ pour /ʧ/).
Les classes dérivées automatiquement correspondent en général à des unités linguistiques identiques ou similaires. Il convient néanmoins de mentionner que 12 d'entre elles sont des paires de phonème d'une même langue : par exemple, le /z/ et le /v/ anglais. Ce cas peut être expliqué par le fait que l'essentiel de l'énergie du [z] est concentrée au dessus de 4 kHz. Par conséquent, le filtrage effectué par la bande téléphonique nous prive d'une grande part de l'information. Inversement, les /u/, /w/, /j/, /l/, /g/, ainsi que les fricatives sonores (notamment /ʒ/) ne sont pas tous rassemblés.

5. CONCLUSION

Nous avons appliqué une approche de classement automatique aux phonèmes de six langues (français, anglais, allemand, espagnol, italien et portugais). Cette méthode est fondée sur un algorithme de classement hiérarchique, qui seul permet un suivi de l'évolution d'un groupement agglomératif. À partir d'une initialisation avec 235 phonèmes, nous avons itéré l'algorithme tout en analysant la pertinence linguistique des groupes obtenus à chaque étape. Nous avons retenu le nombre de 90 classes, afin de ne pas perdre trop d'information phonématique. Ces classes correspondent en majeure partie à une réalité linguistique.
Nous avons aussi analysé les résultats du classement dans les niveaux les plus bas de l'agglomération hiérarchique - où l'on retrouve la dichotomie voyelle-consonne. Ceci est un autre facteur qui démontre l'efficacité de cette méthode. Nous devons aussi noter la capacité de cette approche à traiter des données de différentes origines. De l'analyse des résultats obtenus à chaque itération, nous pouvons observer qu'il n'y a pas eu un classement des phonèmes par corpus, ni un bon classement pour les phonèmes de certaines langues et mauvais pour les phonèmes des autres. Ceci semble démontrer la robustesse de la mesure de

[3] Certains phonèmes en effet, comme les géminées de l'italien, ont peu de représentants. Pour ces phonèmes rares, la validité des modèles statistiques peut être remise en question.

similitude, face aux variations acoustiques des conditions d'enregistrement, et au contenu linguistique des différents corpus d'apprentissage. De fait, la fonction de densité de probabilité correspondant aux vecteurs acoustiques du phonème φ_j, dans l'expression de la mesure de similitude, effectue une normalisation sur toutes les données acoustiques. Cette approche a été appliquée dans le cadre de l'identification automatique de 4 langues (français, anglais, allemand, espagnol). Les résultats obtenus pour le calcul des classes de phonèmes de six langues permettent l'extension du décodeur, de quatre à six langues. On peut même envisager d'utiliser les modèles acoustiques correspondant aux classes de phonèmes des langues connues, pour la reconnaissance phonétique d'une nouvelle langue, dont on dispose uniquement de corpus de parole sans transcription phonétique.

Quantifier la proximité entre les phonèmes et les allophones d'une même langue peut contribuer à la définition d'un ensemble économique de phonèmes. Regrouper les phonèmes de plusieurs langues trouve enfin des applications en synthèse vocale multilingue, et en phonétique descriptive, didactique ou corrective.

Table 2: regroupements d'au moins deux phonèmes pour 90 classes, à travers les 6 langues étudiées (français, anglais, allemand, espagnol, italien, portugais) - une classe par ligne.

Fr.	*An.*	*Al.*	*Es.*	*It.*	*Po.*
i	i	i	i	i	i
e	-	e: ɪ eɪ	-	-	-
ɛ	-	ɛ: ɛ	e	e	ɛ
y	-	y	-	-	-
œ ø	-	-	-	-	-
ə	-	ə	-	-	-
a	æ ɑ: aʊ	a ɑ	a	a	a
ɔ	ɒ	ɔ	o	ɔ o	ɔ
o	ɔ:	o	-	-	o
u	-	-	u	u	u w
-	-	-	-	-	e ẽ
-	-	-	-	-	ĩ j̃
ɑ	-	-	-	-	ɐ
ɔ	-	-	-	-	ũ õ
-	ɜˑ ə	-	-	-	ɐ
-	ɪ e	-	-	-	-
-	ɛ ɛə	-	-	-	-
-	aI	aI	-	-	-
-	-	ɤ əʁ	-	-	-
-	-	u ʊ	-	-	-
w	-	-	-	w	-
j	-	-	-	j	j
-	-	aʊ	-	-	ɨ
s	s	s	s	s ss	s
z	-	-	-	z	z
ʃ ʒ	ʃ	ʃ	-	ʃ ʃʃ	ʒ
f	f	f	f θ	f ff	f
v	-	-	-	v	v
n	n	n ən	n	n	-
-	-	-	-	nn	n
m	m	m	m	m mm	m
ɲ	-	-	ɲ	ɲ ɲ	ɲ
l	-	l	l	l ll	l
ʁ	-	ʁ	-	-	-
p	p	p	p	p pp	p
t	t	t	t	t tt	t
k	k	k	k	k kk	k
g	g	g	-	-	-
-	b	b	-	-	-
-	d	d	-	-	-
-	-	-	b	b bb	b
-	-	-	d	d dd	d
b d	-	-	-	-	-
-	-	-	g	g	g
-	ʤ ʧ	-	-	-	-
-	-	-	-	ʤ ʤʤ	-
-	-	-	-	ʧ ʧʧ	-
-	-	-	-	ts tsts	-
-	z v	-	-	-	-
-	ŋ	ŋ	-	-	-
-	l	-	-	-	l
-	h	h	-	-	-
-	-	-	x	-	ʀ
-	-	-	ǰ ʎ	λλ	ʎ
-	-	-	r r̃	r rr	r

REFERENCES

[1] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, "Developments in large Vocabulary Continuous Speech Recognition of German", *ICASSP-96*.

[2] K.M. Berkling, E. Barnard, "Language Identification of six Languages Based on a Common Set of Broad Phonemes", *ICSLP-94*.

[3] K.M. Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*, PhD thesis, OGI, 1996.

[4] C. Corredor-Ardoy, J.-L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification With Language-Independent Acoustic Models", *Eurospeech-97*.

[5] P. Delattre, *Comparing the phonetic features of English, German, Spanish and French*, Julius Gross Verlag, Heidelberg, 1965.

[6] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, 1973.

[7] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-independent Continuous Speech Dictation", *Speech Communication 15*, pp. 21-37.

[8] D. Gibbon, R. Moore, R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1997.

[9] A.G. Haudricourt, J.M.C. Thomas, *La notation des langues. Phonétique et phonologie*, Imprimerie de l'Institut Géographique National, Paris, 1967.

[10] B.H. Juang, L.R. Rabiner, "A Probabilistic Measure for Hidden Markov Models", *AT&T Technical Journal* 64(2), 1985.

[11] J. Köhler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", *ICSLP-96*.

[12] B. Mak, E. Barnard, "Phone Clustering using the Bhattacharyya Distance", *ICSLP-96*.

[13] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, "The OGI multi-language telephone speech corpus", *ICSLP-92*.

[14] J.G. Proakis, *Digital Communications*, McGraw-Hill Series in Electrical Engineering, 1989.

[15] G.K. Pullum & W.A. Ladusaw, *Phonetic Symbol Guide*, The University of Chicago Press, Chicago and London, 1996.

[16] A. Quilis et J.A. Fernández, 1992, *Curso de fonética y fonología españolas*, Consejo Superior de Investigaciones Científicas, Madrid, 1992.

[17] N. Vallée, L.J. Boë, C. Abry, J.L. Schwartz, A. Berrah, "La matérialité des structures sonores du langage", *JEP-96*.

[18] S.J. Young, P.C. Woodland, "The use of state tying in continuous speech recognition", *Eurospeech-93*.

Production de phones "Nouveaux" et "Similaires" en L2

Hyeon-Zoo KIM

Department of French Language & Literature of Dankook University
29, Anseo-dong Cheonan Choong-nam, South Korea
Tel.: ++343 44 5161, ++343 68 2450 - Fax.: ++343 44 5161
E-mail : hyeonzoo@ppp.kornet.nm.kr

Abstract

A series of experiments are presented, in which identical stimuli are used to examine similar cue-trading relations in the perception of voicing contrast in word stops in Korean and French.

Do Koreans who speak french use the Korean system when uttering french phonems ? Or are they mimicking the French system in the stressed intervocalic position?

Predicted cross-linguistic differences are found in the basis category boundary and in the case of cue trading between VOT and aspiration.

"New" L2 phones have no counterpart in L1 and so, by definition, differ acoustically from phones found in L1. "Similar" L2 phones, on the other hand, differ systematically from an easily identifiable counterpart in L1.

1. Introduction

Dans cette communication, nous allons tenter de définir le rôle du VOICE ONSET TIME (VOT) dans la distinction des trois séries d'occlusives coréennes (faibles: [p t c k], aspirées: [p^h t^h c^h k^h], glottalisées : [p' t' c' k'] et des occlusives sourdes françaises [p t k] à partir de nos propres résultats expérimentaux.

Nous adoptons le même point de vue qu'ABRAMSON (Abr77) & LISKER (Lis78) qui affirment que l'indice de VOT n'est pas une mesure purement acoustique. L'explosion est une conséquence acoustique des mécanismes sous-jacents du larynx et de la cavité supraglottique. Quel que soit le degré d'importance de son rôle dans la distinction des sourdes et des sonores, le VOT mesure le degré de synchronisme du contrôle laryngal et des opérations supraglottiques pendant la production des consonnes occlusives.

Cette étude nous a fourni l'occasion d'examiner la plupart des théories émises jusqu'à cette date à propos du VOT coréen et du VOT français, s'agissant des consonnes que nous avons retenues, à savoir les occlusives qui, en français et en coréen, présentent des différences assez importantes.

Les phones "nouveaux" dans L2 n'ont aucun équivalent dans L1, et donc, par définition sont différents d'un point de vue acoustique. Les phones "similaires" dans L2, en revanche, sont différents systématiquement d'un équivalent facilement identifiable dans L1.

Les questions qui se posent sont les suivantes : Est-ce que les Coréens parlant français utilisent le système coréen quand ils produisent les phonèmes français ou Est-ce qu'ils imitent le système français en position accentuée intervocalique?

2. Préliminaires

2.1. Méthode

Nous exploiterons, dans le cadre de cette étude, les données obtenues à partir de documents oscillographiques mesurés avec le système "Audiomedia" sur ordinateur. "Audiomedia" enregistre directement sur le disque dur, à partir duquel se fait également la lecture. Les processeurs de la carte "Audiomedia" améliorent énormément les capacités audios de l'ordinateur, tout en lui permettant de gérer les tâches exigeantes en temps de cacul.

A partir des données fournies par cette technique, nous avons retenu les paramètres suivants :
- VOT
- Durée totale de la consonne
- Durée de la voyelle subséquente

2.2. Choix du corpus et des sujets

Nous avons choisi les consonnes occlusives en position accentuée intervocalique qui est, en effet, la plus favorable à la comparaison toutes choses étant égales par ailleurs.

Nous avons constitué un corpus, dont les items sont du type V1CV2 (Voyelle-Consonne-Voyelle). Pour étudier les consonnes occlusives, nous nous appuyons sur un premier corpus constitué de 27 phrases en français pour à la fois les trois locuteurs français et les trois locuteurs coréens, ceci afin de bien comparer ces deux groupes. Le deuxième corpus est constitué de 48 phrases en coréen pour les Coréens uniquement. Les phrases contiennent entre 4 et 7 syllabes en coréen et en français et elles sont comparables de par leurs structures syllabiques.

Etant donné le but de notre travail qui consiste à étudier plus spécialement le VOT de consonnes occlusives, nous avons surtout voulu placer, dans les phrases et expressions, les articulations consonantiques dans les principales positions où elles peuvent figurer en coréen et en français par rapport à l'accent et à l'entourage phonique

Les locuteurs français sont de langue maternelle française ; quant aux locuteurs coréens, ils sont de langue maternelle coréenne, étudient le français depuis une dizaine d'années et se trouvent actuellement - janvier 98 - en France.

3. Résultats

3.1. VOT

Comme on le voit dans le tableau 1, les trois groupes (les locuteurs coréens, les locuteurs français et les locuteurs coréens parlant français) ne se superposent pas et l'on relève des différences notables. Chez les Coréens, le VOT des alvéodentales surtout est plus bref que celui des autres consonnes. Le VOT est un indice qui permet de différencier nettement les trois catégories d'occlusives coréennes sourdes. Il en résulte que l'aspiration est l'indice essentiel pour la différenciation entre consonne aspirée et glottalisée. En revanche, c'est la sonorité qui est l'indice de la différenciation entre consonne glottalisée et faible puisque le VOT est égal à O.

Chez les Français, le VOT des vélaires est plus long que celui des dentales et des labiales. Le degré du VOT est le suivant : k > t > p

Nous devons souligner l'importance relative du VOT en particulier pour les consonnes vélaires qui vont jusqu'à 33 ms en moyenne en français. Le VOT français est ainsi supérieur à celui des consonnes glottalisées du coréen sauf dans le cas des palatales coréennes.

Au vu des résultats, la valeur du VOT pour les locuteurs coréens parlant français est moins élevée que celle du VOT pour les locuteurs français et elle correspond à celle des résultats obtenus pour les consonnes occlusives glottalisées du coréen. Pour les valeurs du VOT des Coréens parlant français, on constate une sorte de neutralisation qui fait que les valeurs relevées se situent dans une position intermédiaire entre celles qui caractérisent les aspirées du coréen et celles qui caractérisent les faibles (absence de VOT). Est-ce à dire que ces occlusives sont glottalisées dans la mesure où les valeurs sont très proches? Cela ne semble pas être le cas puisqu'il y a une dispersion plus importante dans le résultat des occlusives dites par les Coréens parlant français par rapport aux glottalisées coréennes qui sont nettement moins dispersées.

3.1. Durée de la consonne

Les durées totales des consonnes aspirées sont plus longues que celles des consonnes glottalisées à cause de leur forte aspiration. Surtout chez les locuteurs coréens, les vélaires aspirées sont beaucoup plus longues que les vélaires glottalisées. Donc, la durée totale est un indice qui permet de différencier les occlusives aspirées et les occlusives glottalisées. Si on admet qu'il y a un rapport entre la durée et la force articulatoire, nous pouvons dire qu'en coréen, les glottalisées, de même que les occlusives aspirées, appartiennent aux catégories des consonnes fortes par opposition aux occlusives faibles dont la durée est nettement plus réduite. Aussi les occlusives faibles [p t k] coréennes ont-elles une durée plus brève qu'en français. Nous remarquons que la réalisation sur le plan de la durée des occlusives dites par les Coréens parlant français est presque toujours plus longue que celle des aspirées et des glottalisées coréennes et a fortiori plus longue que celle des occlusives du français.

Cette particularité ne serait-elle pas due au fait que les locuteurs coréens parlant français auraient mis plus d'énergie et d'attention à la prononciation des phrases et ne serait-elle pas à considérer comme une sorte d'hyper-correction due aux circonstances particulières de l'enregistrement ?

Table 1. VOT, Durée de la consonne et Durée de la voyelle subséquente [les trois groupes : coréen (C), français (F), français parlé par des Coréens (CF)]

	VOT			
	Bila.	Alvé.	Pala.	Vél.
Fai. (C) 0	0	0	0	
Aspi. (C)	60	52	69	88
Glottal. (C)	14	9	34	18
Sourde (F)	15	23	-	33
Sourde (CF)	11	10	-	19

	Consonnes			
	Bila.	Alvé.	Pala.	Vél.
Fai. (C) 57	48	75	62	
Aspi. (C)	169	163	187	175
Glottal. (C)	165	152	178	157
Sourde (F)	146	155	-	145
Sourde (CF)	198	187	-	192

	Voyelle subséquente			
	Bila.	Alvé.	Pala.	Vél.
Fai. (C) 134	134	133	162	
Aspi. (C)	100	114	118	85
Glottal. (C)	135	121	128	123
Sourde (F)	106	94	-	100
Sourde (CF)	118	96	-	99

3.3. Durée de la voyelle subséquente

En français et dans beaucoup de langues, la durée de la voyelle devant les consonnes sonores est plus longue que devant les consonnes sourdes (Tho66, Hou61). Pourtant cet indice ne paraît pas applicable à toutes les langues (Zim58, Kim94). Nous avons travaillé sur cet indice, mais il n'a pas apporté de renseignements valables concernant la nature des occlusives coréennes (Kim95, 96). Alors, pour cette recherche, nous avons décidé d'examiner la durée de la voyelle subséquente.

En ce qui concerne la comparaison de la voyelle subséquente dans les trois groupes étudiés, quel que soit

le lieu articulatoire de la consonne précédente, nous observons que la durée de la voyelle succédant aux consonnes aspirées en coréen, est plus brève que celle des voyelles succédant aux consonnes faibles ou glottalisées. La durée vocalique varie en fonction de la consonne : d'après les observations que nous avons faites, en coréen, la durée de la voyelle succédant à la consonne aspirée [k^h] est beaucoup plus brève et celle de la voyelle succédant à la consonne vélaire /k/ est beaucoup plus longue que celle de la voyelle succédant aux bilabiales, alvéodentales et palatales. Les voyelles succédant aux consonnes faibles [p t c k] (en coréen) ont une durée plus longue que les autres voyelles en coréen et en français. Et chez les Coréens parlant français, la durée des voyelles est proche de celle des Français.

Il faut souligner que dans les occlusives faibles coréennes [p t c k] en position intervocalique, c'est toujours l'occlusive sonorisée qui a la durée totale la plus restreinte. Cependant, le résultat est inverse pour les voyelles subséquentes, qui ont une valeur supérieure aux voyelles. Les valeurs des voyelles subséquentes sont influencées par les consonnes précédentes.

4. Discussion & Conclusion

Comme nous l'avons déjà mentionné dans les résultats, nous avons souvent observé que les occlusives faibles [p t c k] en coréen peuvent phonétiquement être réalisées en sonores en position intervocalique.

Les travaux sur la production de l'occlusive en position initiale chez les bilingues anglais/espagnol (Wil77) ont montré d'une part que les résultats de la production de l'occlusive pour les bilingues sont conformes aux résultats pour les monolingues et, d'autre part, une tendance pour les bilingues au prévoisement [b] en anglais. Dans un travail de Flege (Fle91), il apparaît une alternance (code switching) dans la production de toutes les catégories de locuteurs bilingues, bien que ceux-ci ne puissent exprimer complètement les valeurs monolingues dans leur langue secondaire. Le changement, dans la production du VOT de bilingues anglais/espagnol, est plus important que celui observé par Caramazza dans le cas des bilingues français/anglais (Car74). Il y a aussi une tendance très forte, comme dans la recherche de Williams (Wil77), au prévoisement [b] de bilingues en anglais.

Caramazza & Yoni-Komshian (Car74) ont conclu :
"...VOT is a sufficient phonological cue for the distinction of the homorganic stop consonant pairs in French (France). We have also proposed an explanation for the observed difference between French and Canadian French based on a linguistic change hypothesis," (p. 245).

Ces diverses études ont montré que beaucoup d'adultes qui apprennent l'anglais en L2 produisent les [p t k] en anglais avec des valeurs du VOT significativement plus brèves que celles des Anglais monolingues mais, cependant, avec des valeurs du VOT plus longues que les sujets d'origine monolingue de L1 (Fle81, Por80, Fle87).

Dans un article de Flege (Fle87), il apparaît que les sujets les plus avancés dans leur maîtrise de l'anglais ont produit un [t] néerlandais avec des valeurs du VOT plus brèves que les sujets moins avancés; ce qui suggère qu'ils ont créé une nouvelle catégorie pour le [t] anglais.

Selon quelques travaux antérieurs (Kim94, 95) sur la production de l'occlusive en position intervocalique chez les bilingues coréen/français, nous avons conclu que les Coréens parlant français réalisent une autre forme d'occlusive: car les valeurs du VOT des occlusives sourdes [p t k] produites par les sujets coréens parlant français (et maîtrisant la langue) sont proches des valeurs du VOT du français. La durée des occlusives chez les Coréens parlant français est toujours plus longue que celle des aspirées et des glottalisées coréennes, et plus longue que celle des occlusives du français.

Les résultats actuels montrent qu'en coréen, le VOT est parmi les indices qui permettent de distinguer les trois séries d'occlusives sourdes coréennes, et que le VOT est le plus important dans la distinction entre les occlusives aspirées et glottalisées en position intervocalique. Le VOT du français est supérieur à celui des consonnes glottalisées du coréen. Chez les Coréens parlant français, la valeur du VOT n'est pas très élevée et elle ne correspond pas à celle des résultats obtenus pour les locuteurs français. Mais elle correspond sensiblement à la valeur obtenue pour les consonnes occlusives glottalisées du coréen. Il semblerait toutefois, d'après l'examen de la durée totale, que ni la consonne glottalisée du coréen, ni la consonne sourde du français ne soient utilisées par les Coréens parlant français et, qu'ils n'utilisent pas non plus la consonne aspirée au regard du VOT, ce qui nous a conforté dans l'idée de l'emploi d'une catégorie nouvelle d'occlusive sourde.

Ceci est aussi en accord avec les résultats obtenus par Flege (Fle87, 95, 96) dans ses études concernant l'apprentissage.

Flege (Fle95) indique que:
"...adult learners of a second language may establish new perceptual phonetic categories for phonetic segments not found in the native language."

Nous pouvons dire que les adultes sont capables de créer des phones nouveaux dans une langue seconde, et de modifier leurs modèles articulatoires préalablement établis lorsqu'ils produisent des phones qui leur paraissent identiques à ceux de la langue seconde. Par conséquent, il apparaît que sur le plan acoustique, les sujets coréens parlant français ont tendance à mettre dans la même catégorie des phones différents dans leur langue maternelle et dans leur langue seconde. Finalement, cela pourrait leur permettre de réaliser des phones non pas identiques mais véritablement nouveaux. Nous pouvons donc conclure que les Coréens parlant français réalisent une autre forme d'occlusive. Il s'agit là d'une nouvelle catégorie d'occlusives sourdes.

L'étude de l'interférence linguistique chez les bilingues, de même que la didactique, pourraient-elles éventuellement utiliser et approfondir le concept de catégorie phonique nouvelle ?

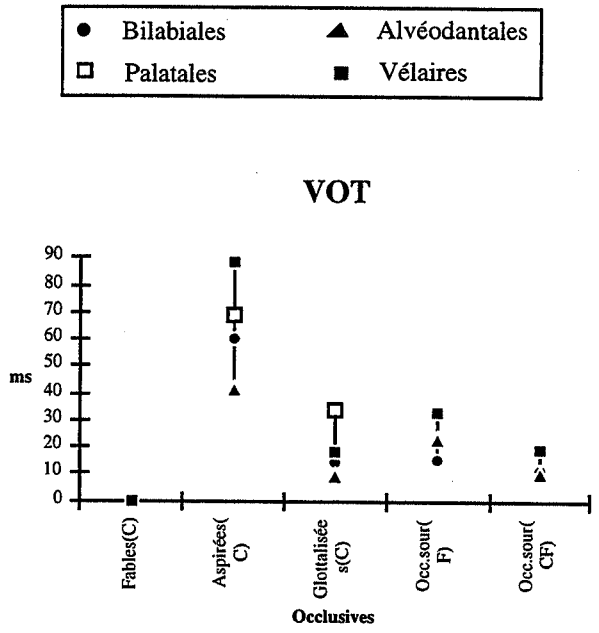


Figure 1. VOT

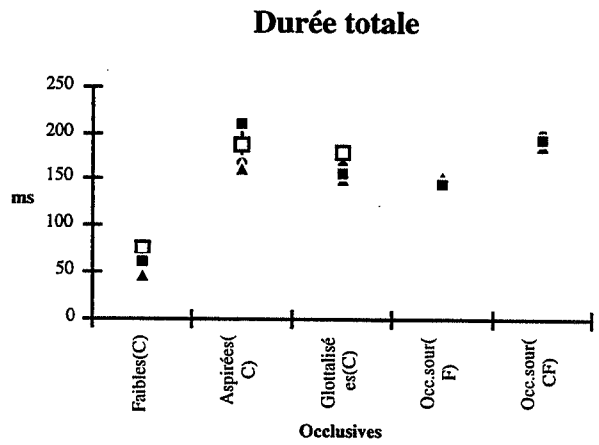


Figure 2. Durée totale de la consonne

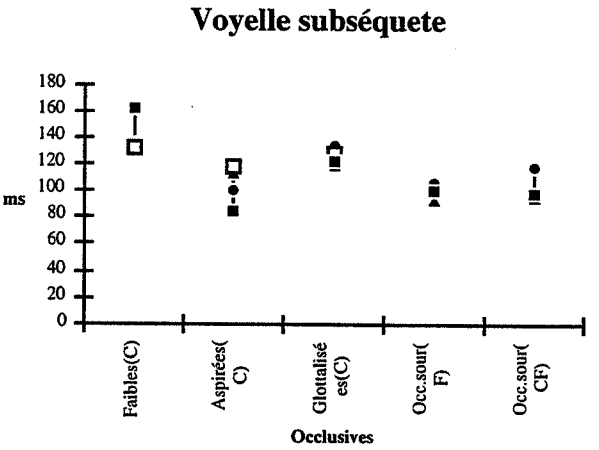


Figure 3. Durée de la voyelle subséquente

4. Bibliographie

[Abr77] Abramson A.S. (1977), "Laryngeal timing in consonant distinction," Phonetica, N° 34, pp. 245-303.

[Lis78] Lisker L. (1978), "In qualified defense of VOT," Language and Speech, N° 21, pp. 375-383.

[Tho66] Thorsen O.M. (1966), "Voice assimilation of stop consonants and fricatives in French and its relation to sound duration and intraoral air-pressure", ARIPUC, Vol. 1, pp. 67-76.

[Hou61] House A.S. (1961), "On vowel duration in English", Journal Acoust. Soc. Am., N° 33, pp. 1174-1178.

[Zim58] Zimmermann S.A. & Sapon S.M. (1958), "Note on vowel duration seen cross linguistically", Journal Acoust. Soc. Am., N° 30, Vol 2, pp. 152-153.

[Kim94] Kim H.Z. (1994), "Contribution à une étude comparative des occlusives du coréen et du français", TIPS, N° 24, pp. 39-89.

[Kim95] Kim H.Z. (1995), "French and Korean Plosives : a Comparative Analysis", ICPhS 95, Vol. 4, pp. 176-179.

[Kim96] Kim H.Z. (1996), "Quelques aspects acoustiques de la production des occlusives du coréen et du français", JEP 96, N° 21, pp. 159-162.

[Wil77] William S.L. (1977), "The voicing contrast in Spanish", Journal of Phonetics, Vol 5, pp. 169-184.

[Fle91] Flege J. (1991), "Age of learning affects the authenticity of voice-onset time (V.O.T.) in stop consonants produced in a second language", Journal Acoust. Soc.Am., N° 89, Vol 1, pp. 395-411.

[Car74] Caramazza A. & Yeni-Komshian G.H. (1974), "Voice onset time in two French dialects", Journal of Phonetics, Vol 2, pp. 239-245.

[Fle81] Flege J. & Port R. (1981), "Cross-language phonetic interference : Arabic to English", Language and Speech, N° 24, pp. 125-148.

[Por80] Port R.F. & Mitler F.M. (1980), "Segmental features and implementation in acquisition of English by Arabic speakers", Journal of Phonetics, N° 11, pp. 219-229.

[Fle87] Flege J. (1987), "The production of "new" and "similar" phones in a foreign language : evidence for their effect of equivalence classification", Journal of Phonetics, N° 15, pp. 47-65.

[Fle95] Flege J & Schmidt A.M. (1995), "Native speakers of Spanish show rate-dependent processing of English stop consonants", Phonetica, N° 52, pp. 90-111.

[Fle96] Flege J., Schmidt A.M. & Wharton G. (1996), "Age of learning affects rate-dependent processing of stops in a second language", Phonetica, N° 53, pp. 143-161.

La partition de l'espace vocalique en berbère rifain

Naïma Louali & Gilbert Puech

UMR Dynamique du Langage
MRASH 14 avenue Berthelot 69363 Lyon Cedex 07
Tel +33 (0)4 72 72 64 93 / Fax : +33 (0)4 72 72 65 90
e-mail : nlouali@mrash.fr / puech@univ-lyon2.fr

Abstract

In Berber, as in Arabic, an emphasis prosody triggered by a pharyngealized coronal consonant covers all segments in the word. The main acoustic cue to discriminate pharyngealized vs non pharyngealized vowels is a lowered F2 value with incident effects on F1.

In Rifian Berber, an etymological *r* may not maintain in coda position and merges with the vocalic nucleus. Sequences *ir and *ur are changed into *ɛæ* and *ɔɑ* respectively. While /a/ is realized *ɛ* or *ʌ* in emphatic context, the sequence *ar yields *æ* or *ɑ* respectively.

The goal of this paper is to demonstrate that the vowel space in Rifian Berber is divided along the F1 axis for vowels, the lower part keeping track of a former *r*, and along the F2 axis for emphatic vs non emphatic vowel discrimination.

1. Introduction

Les parlers du domaine berbère attestent la corrélation d'emphase, également caractéristique du domaine arabe. L'opposition phonologique est portée par les consonnes coronales, qui peuvent être pharyngalisées ou non. Une consonne pharyngalisée initie une emphatisation de tous les segments du mot. Le rifain, dialecte berbère du nord marocain, ne fait pas exception à ce mode de fonctionnement même si l'effet phonétique est moins pervasif qu'en tachelhit ou en tamazight par exemple. Les corrélats acoustiques ont été décrits dans plusieurs études[1]. qui mettent en évidence le rôle discriminant de F2 dans la perception de l'emphase[2].

Dans les parlers rifains, un *r* étymologique[3] ne peut occuper la position de coda ; sa contribution phonologique est alors exprimée par la qualité du noyau vocalique auquel il est associé dans la syllabe. Cette étude montre qu'il en résulte un dédoublement de l'espace vocalique sur F1.

2. Partition de l'espace sur F2

La diffusion du trait de pharyngalisation du segment initiateur vers les segments adjacents entraîne une qualité particulière des voyelles en contexte emphatique comparativement à leur qualité en contexte neutre (non emphatique). Les critères distributionnels déterminent que pour le parler temsamane, seules les obstruantes radicales $\tilde{o}^{\varsigma}$, z^{ς}, t^{ς} et/ou leur contre-partie tendue déclenchent la pharyngalisation et initient une prosodie d'emphase dans les limites du mot.

Nous avons constitué un corpus de 27 termes où les voyelles /i, u, a/ sont placées dans un contexte emphatique ou neutre. Les enregistrements ont été effectués sur DAT en chambre insonorisée et nous avons réalisé les analyses acoustiques avec le logiciel MultiSpeech.

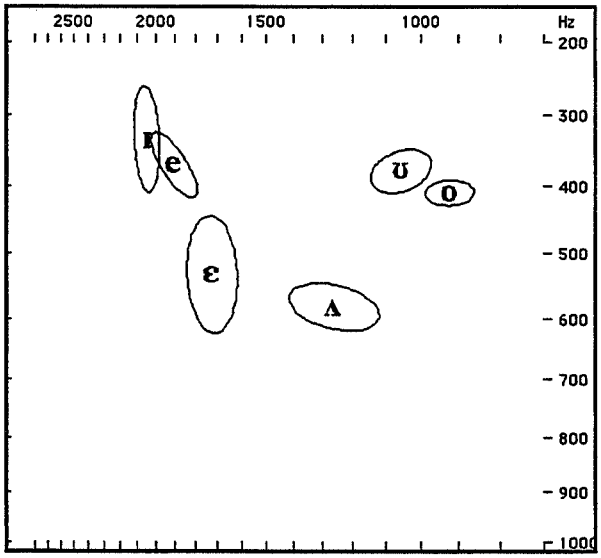


Figure 1. Ellipses de dispersion (15 mesures par ellipse) sur F1/F2 pour les voyelles du rifain de temsamane (locuteur AE) en contexte neutre et emphatique.

La figure 1 montre la formation de couples dans l'espace vocalique suivant que les voyelles sont en contexte neutre ou emphatique :
[ɪ] et [e] pour /i/
[ʊ] et [o] pour /u/
[ɛ] et [ʌ] pour /a/
Le dédoublement s'opère sur l'axe F2 : les voyelles [ɪ], [ʊ] et [ɛ] (contexte neutre) ont un F2 plus élevé que les voyelles [e], [o] et [ʌ] (contexte emphatique). La séparation est particulièrement nette pour la voyelle ouverte. L'incidence sur F1 contribue certainement à la discrimination des allophones mais cette contribution est liée à la partition opérée sur F2.

[1] Notamment : [Ob68], [Gha77], [Ye96] pour l'arabe et [Lou90 et 92] pour le berbère.

[2] [Ye 95] met en évidence la propriété acoustique du début des fréquences de F2 comme seuil critique dans la perception d'une syllabe pharyngalisée vs non pharyngalisée.

[3] Sur la mutation du *l pan-berbère en battue apicale voir [Lou97].

3. Amuïssement du *r*

Les parlers rifains ont connu une évolution particulière de la liquide *r* au sein du domaine berbère. Cette particularité est rendue chez les berbérisants par l'évocation de la "grande instabilité" de ce segment[4]. En position de coda un *r* rhotacise le noyau vocalique, l'allonge ou altère son timbre et sa structure (monophtongue ou diphtongue). Cette répartition peut être envisagée comme gardant la trace de différentes étapes d'une évolution diachronique aboutissant à l'amuïssement du *r* dans cette position. C'est dans le dialecte rifain de temsamane que l'évolution est la plus avancée.

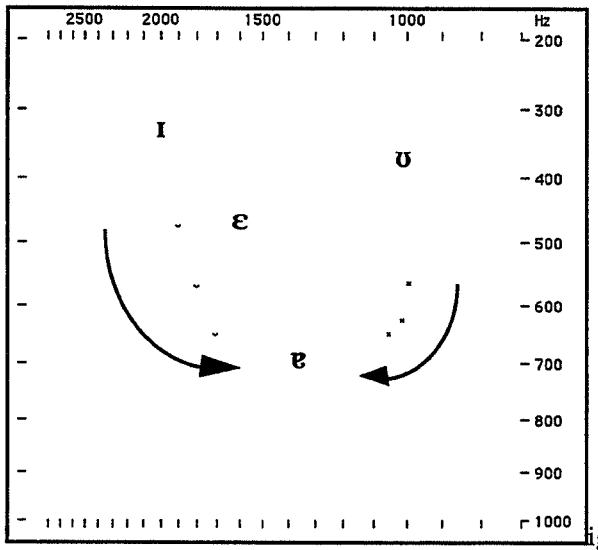
La comparaison avec le dialecte tamazight montre les correspondances suivantes en position coda pour un contexte non emphatique :

æ / ær				
rifain	θæju	*tamazight*	tærju	(ogresse)
	ɛsæðun		æsærdun	(mulet)
ɛæ / ir				
	θɛsɛæθ		tæsirt	(meule)
ɔɑ / ur				
	sɔɑf		surf	(enjambe !)

Ailleurs le *r* se maintient :

intervocalique			
	ɪrɪ	iri	(cou)
	ʊrʊ	uru	(poignée)
appui d'attaque syllabique			
	ɪfrɪ	ifri	(grotte)

La réalisation de la diphtongue *ɛæ* correspondant à /ir/ et celle de son homologue *ɔɑ* correspondant à /ur/ pointent vers une même voyelle ouverte :



ig ure 2. Trajectoire des diphtongues *ɛæ* (ir) et *ɔɑ* (ur) du locuteur AE de temsamane.

La réalisation de la monophtongue correspondant à /ar/ est *æ*. Alors que le timbre des voyelles fermées n'est pas affecté lorsque le *r* est maintenu, la voyelle ouverte se réalise également *æ* quand elle est suivie d'un r maintenu à l'intervocalique:

	ærɪ	(écris)
cf. supra	ɛsæðun	(mulet)

Cette monophtongue, dans le parler considéré, a une durée du même ordre que les réalisations [ɛ] de /a/ . Comme [Dell93] et contrairement à [Chta94] nous concluons à l'absence d'allongement compensatoire pour ce dialecte.

4. Partition de l'espace sur F1

La réalisation ouverte *æ* se substitue à *ɛ* en cas de fusion avec un *r*. Nous avons voulu examiner aussi l'effet de la pharyngalisation. Pour évaluer le couplage des deux processus, nous avons retenu les termes suivants :

/azu/	[ɛzu]	(dépèce !
/arzu/	[æzu]	(cherche !)
/azˤur/	[ʌzˤɔɑ]	(nerf)
/arzˤəm/	[ɑzˤʌm]	(ouvre !)

La répartition de ces différentes réalisations du /a/ dans l'espace acoustique F1 / F2 (figure 3) met en évidence deux séries de couples :

influence du r :	ɛ / æ
	ʌ / ɑ
pharyngalisation :	ɛ / ʌ
	æ / ɑ

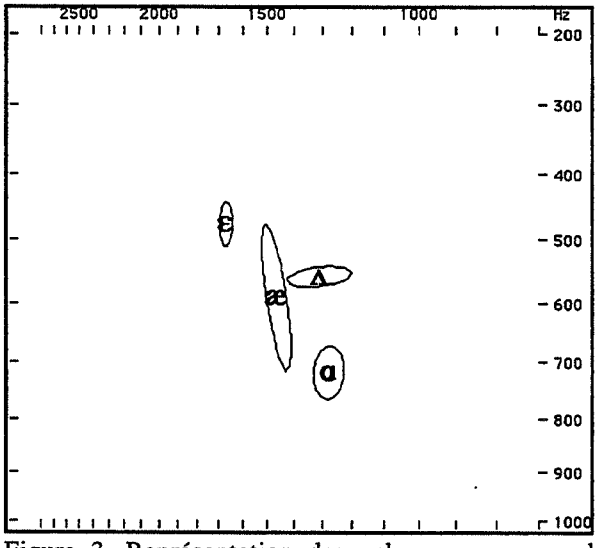


Figure 3. Représentation des valeurs moyennes des différentes réalisations du /a/.

Les deux premiers couples montrent une partition de l'espace vocalique sur F1 tandis que les deux seconds renvoient à une partition sur F2.

5. Influence ouvrante du r ?

L'amuïssement du *r* abaisse-t-il la réalisation de la voyelle ouverte ou prévient-il sa fermeture ? Pour répondre à cette question nous avons étudié le positionnement des voyelles

[4] Pour la description d'autres parlers rifains, cf. par exemple : [Bia17], [Ca87], [Ch79], [El A89], [Ju26], [Laf97], [Re32], [Tan91].

d'un locuteur de tachelhit, dialecte où le *r* se maintient en toute position. La comparaison entre les figures 3 et 4 montrent que les réalisations *æ* et *ɑ* de /a/ en tachelhit correspondent en rifain aux réalisations sous influence d'un *r* amuï. En compensation le rifain ferme les réalisations de /a/ sans influence de *r* par abaissement de la valeur de F1.

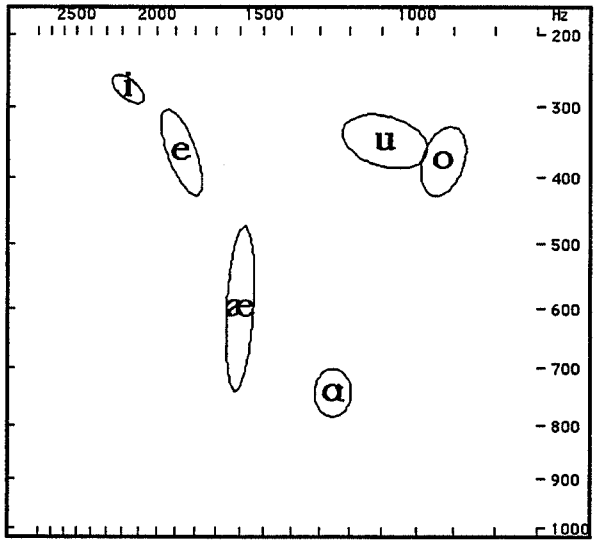


Figure 4. Ellipses de dispersion sur F1/F2 pour les voyelles de Azeni (tachelhit) du locuteur ME en contexte neutre et emphatique.

Conclusion

Le rifain, comme les autres dialectes du berbère, opère une discrimination des réalisations emphatiques vs non emphatiques des voyelles sur l'axe F2. Par suite de l'amuïssement de *r* en position de coda, son originalité est d'utiliser aussi l'axe F1 pour discriminer les réalisations vocaliques. En contexte non emphatique, les trois voyelles du système sont typiquement réalisées [ɪ], [ʊ] et [ɛ] ; lorsque ces voyelles étaient suivies de *r*, les réflexes sont respectivement [ɛæ], [ɔɑ] et [æ]. La comparaison des deux ensembles montre une discrimination sur l'axe F1. La combinaison de la corrélation d'emphase et de l'abaissement compensatoire dû à l'amuïssement du *r* a pour effet d'opérer une double partition fonctionnelle de l'espace vocalique en rifain.

Bibliographie

[Bia17] Biarnay S. (1917), *Étude sur les dialectes berbères du Rif*, Paris : Ernest Leroux.

[Ca87] Cadi, K. (1987), *Système verbal rifain. Forme et sens*. Paris : SELAF.

[Cha79] Chami, M., (1979), *Un parler Amazigh du Rif Marocain. Approche Phonologique et Morphologique*, Thèse de 3e cycle : Université de Paris V.

[Chta94] Chtatou M. (1994), "La représentation vocalique dans les dialectes berbères du Rif", *Études et Documents Berbères*, n° 11, pp. 177-195.

[Dell93] Dell, F. & O. Tangi (1993), "On the vocalisation of /r/ in ath-Sidhar Rifain Berber", *Linguistica communicatio* . Fès (Maroc), 5/1-2, pp. 211-224.

[El A89] El Aïssati, A., (1989), *A Study of the Phonotactics of Asht Touzine Tarifit Dialect*, Thèse de 3e cycle : Faculté des lettres, Université Mohamed V, Rabat.

[Gha77] Ghazali, S. (1977), *Back Consonants and Backing Coarticulation in Arabic*. Doctoral Dissertion, University of Texas at Austin.

[Ju26] Justinard Cmdt (1926), *Manuel de Berbère Marocain : dialecte rifain*, Paris : Paul Geuthner.

[Lad 96] Ladefoged P. & Maddieson I. (1996), *The Sounds of the World's Languages*, Blackwell Publishers Ltd.

[Laf97] Lafkioui M. (synthèse des travaux et conclusions élaborée par —) (1997), veau doctorat : Université de Paris VIII.*Propositions pour la notation usuelle à base latine du rifain*, Inalco, Paris.

[Lin85] Lindau M. (1985), "The story of r" In *Phonetic Linguistics*, ed. by V. A. Fromkin, Academic Press, Orlando, FL : 157-68.

[Lou90] Louali N. (1990), *L'emphase en berbère. Étude phonétique, phonologique et comparative*, Thèse de doctorat : Université Lumière-Lyon 2.

[Lou92] Louali N. (1992), "Le système vocalique touareg", *Pholia*, n° 7, pp. 83-115, Université Lumière-Lyon 2.

[Lou97] Louali N. & G. Puech (1997), communication sur «Le vocalisme en berbère rifain», Journées d'Etudes Linguistiques : *La voyelle dans tous ses états*", Nantes, 5-6 décembre 1997.

[Ob68] Obrecht, D. H. (1968), *Effects of the second formant on the perception of velarization consonants in Arabic*. The Hague, Mouton.

[Re32] Renisio, A. (1932), *Étude sur les dialectes berbères des Beni Iznassen, du Rif et des Senhaje de Srair*, Paris : Ernest Leroux.

[Ta91] Tangi, O., (1991), Aspects de la phonologie d'un parler berbère du Maroc : Ayt-Sidhar (Rif), Thèse du nouveau Doctorat : Université de Paris VIII.

[Ye95] Yeou M. (1995), «Trading relations between cues for the pharyngealized/non-pharyngealized contrast», *Proceedings of the XIII International Congress of Phonetic Sciences* (14-19 Août), Volume 4, pp. 464-467.

[Ye96] Yeou M. (1996), *Une Étude expérimentale des consonnes postérieures et pharyngalisées de l'arabe standard*, Thèse de doctorat : Université de la Sorbonne Nouvelle, Paris III.

La phonétisation des emprunts et de leur paradigme flexionnel

Sannier Frédérique, Belrhali Rabia, Aubergé Véronique

Institut de la Communication Parlée, ESA CNRS 5009
Université Stendhal/INPG, domaine universitaire
38040 Grenoble Cedex
Tél.: 04 76 82 41 17 - Fax: 04 76 82 43 35
e-mail : (sannier,belrhali,auberge)@icp.inpg.fr

ABSTRACT

Our aim is here to give a survey of the phonographical behaviour of the loanwords introduced in the French lexicon, through the observation of the systematic functioning of the French letter-to-phone TOPH system. We thus define sub-systems, isolated into lexicons. The observation of the utterances graphonical functionings, made it possible to delimit classes. The second part of this study more specifically deals with the loanwords inflexion paradigms, for which as well different functionings are drawn.

1. INTRODUCTION

Nous avons entrepris d'établir une description générale des systèmes de transcription graphoniques internes aux lexies du français [Bel95]. Nous avons pour cela adopté une méthodologie inductive basée sur l'observation d'un large échantillon attesté : les formes orthographiques et phonétiques d'un lexique général de la langue, Le Petit Robert 1 [Rob91]. Partant d'hypothèses triviales sur les correspondances graphies-sons impliquées dans la transcription, nous avons fait émerger, par analyse du lexique, des unités minimales de correspondances – les graphones – (pour une définition de cette unité, voir [Aub96]). La taille de ce corpus (59 000 lexies pour [Rob91]) nous permet de prétendre généraliser à la langue les processus mis en évidence dans ce lexique. Ainsi, l'analyse complète de [Rob91] fait apparaître, pour certains graphèmes, des contraintes contextuelles qui délimitent des "zones lexicales", c'est-à-dire de sous-ensembles de lexies pour lesquelles la transcription graphonique suit un système de régularité spécifique. C'est le cas de nombreuses lexies d'emprunt (à travers éventuellement une transcription inter-alphabets ou une alphabétisation) pour lesquelles on peut circonscrire à l'intérieur de [Rob91] des espaces linguistiques dans lesquels le processus de phonétisation résulte de la projection sur le français du système de la langue source. Nous présentons en première partie des sous-systèmes relatifs à quelques langues sources représentées par des lexiques extraits de [Rob91]. Ensuite nous nous intéressons aux phénomènes de phonétisation des formes fléchies des emprunts. L'ensemble de ces descriptions est implémenté dans une grammaire de phonétisation automatique écrite en langage TOPH et a été testé en boucle sur un lexique équivalent à [Rob91].

2. LES EMPRUNTS : DES SYSTÈMES SPÉCIFIQUES

2.1 La difficulté d'identification de l'emprunt

Même s'il peut être ressenti par certains de ses locuteurs comme une menace sur l'identité d'une langue donnée, l'emprunt, tout comme la créativité lexicale peut aussi être perçu comme preuve de dynamicité et d'adaptation de la part de la langue en question. L'emprunt est un phénomène dont toutes les langues en contact font l'expérience, par le media premier que sont ses usagers, et il est à distinguer de l'interférence ou de l'alternance entre les langues [Pop84]. Mais il convient de donner une première définition, inévitablement réductrice : l'emprunt consiste en la translation dans un système linguistique donné de caractéristiques (ici phonétiques) qui sont propres à un autre système linguistique. Les motivations sont diverses : la lexie empruntée peut l'avoir été en même temps que son référent (*tomate*, exemple qui sera développé plus avant) ; la lexie n'a pas d'existence dans la réalité extra-linguistique (*djemââ*, 1870, arabe, réunion de notables qui représentent un douar, en Afrique du Nord) ; à un autre niveau, politique, la lexie est introduite sous la pression d'une communauté linguistique dominante (cela concerne en particulier le langage de l'informatique).

Mais si l'emprunt récent est facilement repérable car encore empreint d'exotisme dans sa prononciation (nous prenons bien sûr le terme "exotisme" dans son acception la plus large : est exotique tout ce qui vient de l'extérieur, de l'étranger), d'autres perdent la trace (morphologique, phonétique...) de leur origine exogène. Reprenons l'exemple de *tomate*. Cette lexie provient du nahuatl, une des langues, idéographique, des aztèques. Elle a donc subi une orthographisation, d'abord en espagnol (passant de *xitomatl* à *jitomatl*), puis une modification de la prononciation, glissant de [tl] à [ta], pour la finale, puis [te], qui est devenu [t] en français.

La morphologie de la lexie peut conduire à des égarements : si *choucroute* est bien le résultat de la composition de *chou* et *croûte*, elle n'en a pas moins été générée par attraction phonétique de l'alsacien *Sauerkraut*. L'assimilation est ici pleinement achevée : les morphologies graphiques et phonétiques appartiennent entièrement au système linguistique du français. Par contre, les lexies du lexique général du français présentant la séquence graphique ‹tch› se prononçant [tʃ] à l'initiale, en revanche, gardent par là même la trace de leur origine étrangère puisqu'aucune

lexie non empruntée ne la présente (*tchin tchin* : 1902, pidgin-english de Canton ; *tchervonetz* : XX°, Russe).

2.2 Le traitement des emprunts dans TOPH

La grammaire de phonétisation écrite en langage TOPH permet un traitement spécifique de la prononciation des emprunts (si tant est que ceux-ci présentent des déviations par rapport à la norme), et ceci par deux moyens : (1) l'écriture de règles individuelles qui témoignent par là même de la singularité de la correspondance graphème-phonème (‹eo› => [i] uniquement dans *boat people*, 1979, anglais) ; (2) des lexiques, qui sont le résultat d'un filtrage systématique de [Rob91] à travers le système général du français. La méthodologie est la suivante : les unités minimales de transcription présentant une déviance par rapport à la grammaire primitive servant de base sont regroupés dans des lexiques, qui sont donc la trace d'un fonctionnement spécifique unifié. Ainsi sont circonscrites les lexies qui, d'une part sont issues d'une même langue source, et qui, d'autre part ont la même prononciation pour chaque unité donnée.

L'ensemble des traces diachroniques que constituent ces règles et ces lexiques est donc l'image de la correspondance graphème-phonème de la langue source translatée sur le système phonétique du français. De la même façon que nous avons proposé des principes d'analyse (unité minimale, attestée diachroniquement) et de fonctionnement (unité partagée par d'autres graphèmes) pour définir les graphones dans le système général [Aub96], il faudrait dans l'idéal appliquer ces principes et définir les graphones à l'intérieur même du système général de la langue source de l'emprunt afin de le récupérer, transféré, dans la langue cible. Puisque nous n'avons pas défini les graphones à l'intérieur des langues source (ce qui se révélerait d'ailleurs très complexe dans le cas d'alphabets différents, sans mentionner les langues sources non alphabétiques) il nous faut appliquer les principes de délimitation du graphone directement sur les lexiques isolés dans [Rob91], faibles échantillons de la langue source.

2.4 Classement phonétique des emprunts

Certaines lexies sont les témoins de ce que nous appellerons des sous-systèmes (*tabac*, issu de l'arouate, langue d'Haïti, *tsibatl*, par l'espagnol *tobaco*) à l'intérieur d'un système de réécriture plus large (*c* en position finale ne se prononce pas *(clerc)*). Ces sous-systèmes de transcriptions propres aux phénomènes des emprunts suivent quatre configurations :

Premier cas L'unité appartient également au français et est prononcé dans la langue cible selon un transfert immédiat *(yacht [jak])*. Les unités sont donc celles du système général du français, ce qui implique logiquement que les lexies ne font pas l'objet d'un traitement spécifique dans la grammaire.

Deuxième cas L'unité appartient également au français mais les cibles sonores changent *(pace maker [pɛsmɛkœʀ])*. Il est à noter que l'anglais a un statut particulier [Cat71].

Troisième cas La cible sonore de l'unité empruntée est une généralisation contextuelle du système général (*bled [blɛd]*, argot militaire de l'arabe algérien, fin XIX°) ou ont été absorbées *(wagon [vagɔ̃], 1826, warranter [vaʀãte], 1874)*.

Quatrième cas L'unité héritée n'appartient pas au français et sa cible sonore résulte d'un transfert sans analogie avec le système français. La consonne [χ], constrictive uvulaire, spécifique des systèmes phonologiques arabe et espagnol, a subi plusieurs adaptations graphiques. Cette consonne se transcrit *kh* (***kh**amsin*), *gh* (*ma**gh**zen*), *ch* (*c**h**amsin*) et *j* (*azulejo*). Quant à sa prononciation dans la langue cible, en l'occurrence le français, elle est devenue [x], constrictive vélaire.

Dans les tableaux récapitulatifs ci-dessous, nous présentons un extrait des graphones pour les langues sources anglais et arabe (issu de [Bel95]).

Tableau 1: graphones des emprunts de l'anglais

aw-->[o] (8)	crawl 1908	eo - [i] (1)	boat people, 1979
a-->[ɛ] (16)	pace maker 1960	a --> [a] (28)	gentleman 1698
a --> [a] (4)	tramway 1818	a -->[œ] (1)	breakfast 1865
ch --> [] (7)	yachting 1859	d --> [w] (2)	midshipman 1785-1867
ea -->[ɛ] (9)	breakfast 1865	ea-->[w] (30)	leadership 1878
e --> [] (14)	briefing 1945	e --> [ɛ] (29)	stencil 1915
ew-->[u] (2)	interview 1883	e --> [œ] (1)	after-shave 1962
g --> [] (9)	copyright 1830	g --> [dʒ] (17)	gentry 1688
i-->[aj] (33)	esquire 1870	i --> [] (7)	business 1861
i --> [œ] (9)	girl XX^{e}	k -->[] (7)	knickers 1863
l --> [əl] (4)	boat people 1979	ng-->[ŋ] (84)	camping 1905
n --> [n] (4)	brainstorming 1950	oo--> [u] (45)	coolie 1575
ou-->[aw] (13)	output 1965	ou --> [ɔ] (1)	dreadnought 1906
o-->[œ] (1)	patchwork 1964	r --> [æʀ] (1)	esquire 1870
s --> [] (3)	schelem 1773	sh-->[ʃ] (35)	brushing 1966
th-->[s] (2)	granny smith 1968	t --> [ti] (2)	tshirt 1950
u -->[œ] (32)	runabout 1934	u --> [u] (17)	bush 1926
u -->[ju] (4)	fuel XX^{e}	u--> [] (1)	building 1895
w -->[] (15)	rewriting 1945	y --> [aj] (5)	skye-terrier 1891

Tableau 2 : polyphones de l'anglais

a --> [e,a]	baby-sitting 1960	a -->[ɔ,ɑ]	talkie-walkie 1954
ch --> [k,ʃ]	schooner 1800	ch --> [ʃ, tʃ]	sandwich 1801
d --> [t,d]	steward 1846	e --> [i,ju]	steward 1846
ee--> [ə,i]	engineering 1953	e --> [ɪ,e]	design 1965
ou -->[aw,u]	round 1816	u --> [y,ju]	tuner 1960
u -->[œ,ɔ]	flush 1933		

Tableau 3 : graphones des emprunts arabes

â (1) --> [ɑ]	djémâa 1870	j (1) --> [j]	kandjar 1617
gh (1) --> [x]	maghzen 1849	j (1) --> [x]	azulejo 1846
kh (1) --> [x]	khamsin XVIII^{e}	s (1) --> [z]	casbah 1830

2.5 Les polyphones

Dans la grammaire est effectué un traitement spécifique des polyphones. Sont proposées pour chaque lexie les différentes prononciations attestées dans [Rob91]. Ainsi *design* se réécrit [dezajn] et [dizajn]. Il arrive souvent qu'une même lexie *(pull-over)* comporte une prononciation régularisée sur le système général du français [pylɔvɛʀ], et une autre plus proche du phonétisme d'origine [pulɔvœʀ].

3. LE SYSTÈME FLEXIONNEL DES EMPRUNTS

3.1 Repères linguistiques

Recentrons notre propos sur un phénomène particulièrement intéressant de l'emprunt, celui de son système flexionnel. Ce champ d'observation mérite un intérêt particulier par au moins un égard : le paradigme flexionnel d'une lexie empruntée témoigne souvent, même si cette affirmation se devra d'être nuancée, du degré d'intégration de celle-ci.

Le paradigme flexionnel semble être un bon moyen d'établir synchroniquement une sorte d'échelle graduée permettant de juger du degré d'intégration de telle ou telle lexie empruntée dans une langue source donnée. Pour reprendre l'exemple de *tomate*, il s'agit d'une lexie féminin singulier qui forme son pluriel sur le schéma le plus récurrent du français : l'adjonction d'un *s* graphique. La prononciation, quant à elle, n'est pas affectée, ce qui s'éloigne du système de la flexion en nombre de l'espagnol, langue qui sonorise les *s* finaux, et à plus forte raison du schéma nahuatl, dont le nombre se forme par l'ajout d'une voyelle qui s'assimile en français au [i], *(tomatli)*. Bien peu de locuteurs du français identifieraient cependant la lexie *tomate* comme étant un emprunt.

On constate plusieurs configurations, qui se doivent d'être décrites en extension. Dans le premier cas la forme fléchie est empruntée en même temps que la forme canonique, et subit un transfert phonologique sur le français *(yacht(s)man)*. Ensuite la forme peut être fléchie selon le système flexionnel du français après transfert phonologique de la forme canonique *(scout)*. Le troisième cas est celui où les attributs flexionnels (principalement la distinction singulier/pluriel) de la forme sont transformés après le transfert *(blinis)*. Enfin, la forme se fléchit selon un système transféré de la langue source, ce qui peut conduire à des formes erronées *(tenniswoman)*.

3.2 La génération automatique des formes fléchies

Deux outils Hypercard [San97] ont été élaborés, qui permettent de générer les formes fléchies de deux lexiques : un lexique de substantifs, pouvant s'employer (ou non) également comme adjectifs, et un lexique d'adjectifs pouvant s'employer ou non comme substantifs (les fonctionnements sont disjoints). Ces lexiques sont extraits d'un lexique plus vaste car contenant également des verbes, le "60 000" de l'ICP. Dans ces générateurs, chaque lexie subit des opérations préalablement mises en évidence dans [Cha91], qui les classe selon la nature du paradigme flexionnel, le genre et le nombre, et la nature de la différence de prononciation entre la forme canonique et la ou les formes fléchies.

Les unités manipulées ne sont pas des unités linguistiques, mais des unités minimales de substitution, bien que celles-ci se recouvrent parfois.

ex : *chasseur, chasseresse* => *ur/resse* (signifie : *resse* se substitue à *ur)* : pas de recoupement de l'unité linguistique et de l'unité de fonctionnement. Le découpage en morphèmes donne *chass-eur* et *chass-eresse*.

carbonaro, carbonari => *o/i* : les unités se recoupent.

Il n'y a pas de module d'analyse morphologique même si les unités substituées sont des indices sur le mode de formation des formes fléchies quand on les rencontre avec une certaine récurrence.

ex : *jazzman, jazzmen* => comportement flexionnel identique à *rugbyman (rugbymen), policeman (policemen)*, et quinze autres lexies présentes dans notre classification, car ces lexies appartiennent au même sous-système. Cet exemple manifeste la récurrence d'un mode de fonctionnement du pluriel anglo-saxon et illustre un phénomène intéressant. Si le locuteur de la langue cible a conscience de l'appartenance de la lexie à une langue source particulière, ici l'anglais, ce locuteur peut déduire le fonctionnement flexionnel d'une lexie canonique morphologiquement identique au moins en partie. Ceci allant jusqu'à produire des formes erronées, comme il sera montré dans la suite de cet exposé.

3.3 Les sous-systèmes

S'il est difficile de trouver dans la littérature des documents sur le transfert morphophonologique des formes fléchies des emprunts, la phonétisation automatique du "60 000" par le phonétiseur TOPH et de ses formes fléchies nominales et adjectivales, a permis de dégager les grands cas de transfert, à la fois graphiques et phonologiques. Elles vont être présentées selon une gradation qui pourrait être celle mentionnée plus haut.

Premier cas La langue cible récupère tout le paradigme flexionnel de la lexie soumise à l'emprunt, au niveau graphique, et subit un transfert phonologique le plus fidèle possible à la langue source. Exemple : la série *yacht(s)man, yacht(s)men, yacht(s)woman, yacht(s)women,* dont la première apparition date de 1859, empruntée à l'anglo-saxon, est passée intégralement au français. La prononciation de la base (le graphème ‹a› ne se prononçant [ɔ] qu'à une fréquence inférieure à 1% sur notre lexique), tout comme celle des morphèmes flexionnels, respecte celle de l'anglais. Tel est aussi le cas de *pirojki*, [piʀɔʃki] qui n'a qu'une entrée masculin pluriel dans [Rob91], mais dont le singulier existe pourtant en français : *pirojok* ([piʀɔʒɔk]).

Deuxième cas On emprunte la graphie de la forme canonique et on y plaque une prononciation et une

flexion françaises : *boy scout [bɔi skaut]* a donné *scout [skut]*, et *girl scout, [gəɹl skaut] scoute [skut]*. Il existe aussi un fonctionnement mixte *(maximum [maksimɔm], maximums [maksimɔm], maxima [maksima])* où la flexion se fait "à la latine" ou "à la française". Il est pertinent ici de reprendre la remarque faite quant à la datation de l'emprunt qui justifierait tel ou tel degré d'intégration : la première apparition en français de *maximum*, d'après [Rob91], date de 1718. Si l'on reste sur les exemples de *yachtman* et *maximum*, la théorie se tient. Mais elle est mise à mal par d'autres exemples, tel que *goy [gɔj]* (de l'hébreu, "chrétien"), qui a les mêmes caractéristiques que le premier cas (pluriel *goym [gɔim]* ou *goyim [gɔjim])* mais dont la datation est bien plus ancienne (XVI^{e} siècle). D'autres facteurs entrent en jeu par conséquent, parmi ceux-ci la fréquence d'utilisation [Pop84] [Hol76] (*maximum*, est entré plus récemment dans le lexique du français, mais a une fréquence d'utilisation plus élevée que *goy*).

Troisième cas Le français ne prend que la forme fléchie de la lexie et la recatégorise en forme canonique. Il perd ainsi une trace de la flexion et superpose sur ce qui est devenu une forme non fléchie son morphème flexionnel. Ex : *blinis [blinis]*, emprunté en 1883, vient du russe *blini ([blini])*, pluriel de *blin ([blin])*. On aboutit ainsi en français à une redondance du nombre. Ce phénomène a été relevé par Haugen [Hau50], qui le nomme "flexion hybride" (p.218). Quant à la sonorisation du *s* final, elle est très certainement analogique.

Quatrième cas Le français ne reprend que la base canonique et plaque un paradigme flexionnel qu'il croit être celui de la langue source. Il en transfère par là-même la prononciation. Ex : *tennisman [tenisman], tenniswoman [teniswuman], tennismen [tenismɛn], tenniswomen [teniswumɛn]*. Ces formes sont des pseudo anglicismes, inventés en 1934 sur le modèle des lexies ayant le même comportement que *sportsman*. L'anglais ne fait pas de distinction de genre, seulement de nombre : *tennis player, tennis players*.

Les quatre grands cas de fonctionnement que nous avons pu dégager présupposent l'autonomie des morphèmes linguistiques que sont la base et la flexion. Contre toute attente, une forme empruntée ne l'est donc pas forcément avec tout son bagage flexionnel.

Les études traitant de l'emprunt soulignent le fait que les adaptations d'éléments étrangers de tout niveau linguistique aux patterns de la langue réceptrice, telles que l'incorporation de suffixes verbaux et nominaux, l'attribution du genre, et la mise en place du paradigme flexionnel, etc. constituent des indications que les formes ont été intégrées dans cette langue.

4. CONCLUSION

L'origine linguistique d'une lexie conditionne fortement sa prononciation. Pour oraliser une lexie d'origine étrangère, il convient d'adapter la prononciation aux règles de phonétisation considérées comme compatible avec l'origine supposée. Mais pour une même langue source, on distingue plusieurs stades d'entrées, correspondant à des usages différents. Les stades contemporains sont représentés dans notre grammaire par un lexique des entrées trouvées dans [Rob91] (qui décrit en extension). De cette observation on peut envisager, comme il est fait dans le système général, étendre ce sous-système à toutes les entrées éventuelles répondant aux mêmes caractéristiques (date, usage...) Ainsi, plutôt que de contraindre les règles de ces sous-systèmes par des lexiques, on pourrait associer un détecteur de ces attributs (définition par intention) en utilisant par exemple une structure cohorte de la langue source comme identification d'origine [Bec96].

BIBLIOGRAPHIE

[Aub96] Aubergé V.& Belrhali R. (1996), La phonétisation automatique du français : émergence de règles ou de lexiques ? Revue LIDIL n°13.

[Bec96] Béchet F. (1996), Intégration de différents niveaux linguistiques pour le traitement des mots hors-dictionnaire dans la conversion graphème-phonème automatique, XXI^{e} JEP, Avignon.

[Bel95] Belrhali R. (1995), Phonétisation automatique d'un lexique général du français : systémique et émergence linguistique. Thèse de l'Université Stendhal, Grenoble 3.

[Cat71] Catach N. (1971), Orthographe et lexicographie, tome 1, éditions Didier, Paris.

[Cat95] Catach N. (1995), Dictionnaire historique de l'orthographe française. Éditions Larousse, Trésors du Français.

[Cha91] Chatti S. (1991), Catégories lexicales du français. T.E.R. de maîtrise, Université Stendhal, Grenoble 3.

[Gui75] Guilbert L. (1975), La créativité lexicale, collection langue et langage, Larousse Université, Paris

[Hau50] Haugen E. (1950), The analysis of linguistic borrowing, "Language" n°26, pp.210-231.

[Hol76] Holden K. (1976), Assimilation rates of borrowings and phonological productivity, "Language" n°52, pp.131-147.

[Pop84] Poplack S., Sankoff D. (1984), Le trajet linguistique et social des emprunts. "Revue québécoise de linguistique" Vol 14, pp.141-186.

[Rob91] Petit Robert 1 (1991), Édition Le Robert.

[Rob96] Robert Étymologique en deux volumes. Éditions Le Robert, Paris.

[San97] Sannier F., Aubergé V., Belrhali R. (1997), La phonétisation des morphèmes flexionnels du français dans le système TOPH. Actes des 1ères JST, AUPELF_UREF Francil, Avignon, pp.463-468

Organisation Temporelle des Voyelles dans les Structures CVCVCV, CVCCVCV et CVCCV de l'Arabe Standard.

A. Amrouche[1], B. Boudraa[1], J.M. Rouvaen[2]

[1]Lab. Signaux et Systèmes, Institut d'Electronique,
USTHB, B.P. 32, Bab Ezzouar, 16 111, Alger, Algeria.
tel: (213)2515575, fax:(213)2515039, e-mail: usthbele1@ist.cerist.dz
[2]Dépt. O.A.E, U.M.R C.N.R.S 9929,
Institut d'Electronique et de Micro-électronique du Nord, UVHC
Le Mont Houy, B.P. 311, 59304, Valenciennes cedex, France.
tel: 33 327141238, fax: 33 327141189, e-mail: rouvaen@ univ-valenciennes.fr

Résumé

Dans cet article, nous nous intéressons à la variation de la durée vocalique dans trois structures usuelles de la langue Arabe.
Après avoir mesuré les durées intrinsèques des voyelles longues et brèves, nous avons étudié les influences des segments adjacents pour déterminer les contextes dominants. Nous montrons alors que la présence de consonnes géminées ou de groupements consonantiques réduit sensiblement la durée de la voyelle qui les précède.

Mots clé: durée vocalique, durée intrinsèque, contexte dominant.

1. Introduction

Notre présente étude est une investigation dans le domaine de l'organisation temporelle qui régit les structures de la langue Arabe. Elle complète une étude analogue menée sur les consonnes de cette langue. L'objectif recherché est l'élaboration d'un modèle prédictif de durées des phonèmes de l'Arabe standard, inspiré du modèle de Klatt.

2. Durées Intrinsèques des Voyelles

2.1 Durées Intrinsèques des Voyelles Brèves

Le contexte environnant pour la mesure de la durée intrinsèque procède toujours d'un choix subjectif. Pour notre étude, nous avons maintenu le contexte occlusif voisé /b/ et /d/ [Kla76], mais celui-ci doit être pris dans des mots respectant un schème de la langue Arabe. La durée intrinsèque des voyelles brèves a été mesurée dans des expressions de type C1V1C2V2C3V3 en position V1. Si pour les voyelles /a/ et /u/ il était facile de trouver des expressions de la langue Arabe respectant les schèmes /façila/ et /fuçila/, pour la voyelle /i/, nous avons utilisé des logatomes de type /fiçula/.

Deux locuteurs, un Algérien et un Syrien, ont répété six fois les mots de la liste suivante; badiça / budiça / biduça; badiza / budiza / biduza; insérés dans des phrases porteuses /qul...fula:n/, dont la traduction veut dire /dis, il a...(fait) untel/. La vitesse d'élocution des deux locuteurs est comprise entre quatre et cinq syllabes/sec. Après les mesures des durées Di des phonèmes, nous avons calculé les paramètres suivants:

-la durée moyenne du segment, soit $D0 = \frac{1}{n} \cdot \sum_{i=1}^{n} Di$

-l'écart maximal de durée soit, $\Delta max = \text{Max} \mid D0 - Di \mid$

-l'écart type σ, tel que $\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} [Di - D0]^2$

A titre comparatif, nos résultats, exprimés par le tableau 1, sont proches de ceux proposés par des auteurs ayant utilisés des phrases porteuses [Ben82, Mit84]. Ghazali et col. [Gha92] proposent une durée de la voyelle /a/ égale à 91 ms pour un locuteur parlant avec une vitesse d'élocution de 4.5 syllabes/sec. (valeur sensiblement équivalente à la notre) et une durée de 54 ms pour cette même voyelle prononcée avec une vitesse d'élocution de 6.6 syllabes/sec. Mrayati [Mra84] situe l'intervalle des durées des voyelles brèves entre 100 et 150 ms. Par contre, El-Ani [Ani70], qui a utilisé des mots isolés, propose des valeurs encore plus élevées, allant jusqu'à 300 ms.

Les durées mesurées font apparaître une nette différence, de l'ordre de 12ms, entre la durée de la voyelle antérieure /a/ et celles des voyelles postérieures /u/ et /i/. Les relations d'ordre DUR /a/ >DUR /u/ et DUR /a/ > DUR /i/ ont été vérifiées sur la première syllabe accentuée. Ces relations sont moins évidentes pour des positions intermédiaires, ou finale, du phonème dans le mot. Pour Benkirane [Ben82], qui s'est intéressé à l'arabe Marocain, cet ordre relève des caractéristiques intrinsèques du locuteur et du contexte.

Tableau 1: Durées intrinsèques (en ms) des voyelles brèves de l'Arabe standard.

	Alg.			Syr.		
phonème	/a/	/u/	/i/	/a/	/u/	/u/
durée moy.D0	97	84	80	91	80	77
durée Dmin	78	71	65	77	73	65
durée Dmax	110	97	97	103	90	84
ecart Δmax	19	13	17	14	10	12
ecart type σ	8.39	8.15	9.55	8.77	4.96	6.81

2.2 Durées Intrinsèques des Voyelles Longues

Les durées intrinsèques des voyelles longues ont été mesurées dans des schèmes de type CVCV:CV ou V: représente la voyelle longue. Les deux locuteurs algérien et syrien ont prononcé six fois les mots, insérés dans des phrases porteuses, de la liste: */bada:da/, /badu:da/, /badi:da/, /ɕaba:da/, /ɕabu:da/, /ɕabi:da/.*

Les durées des voyelles longues, données dans le tableau 2, sont proches de celles de Mrayati [Mra84], dont l'intervalle se situe entre 225 et 250 ms. Elles sont néamoins supérieures à celles trouvées par des auteurs ayant utilisé des contextes semblables [Ben82, Gha92], mais trois fois moins élevées que celles proposées par El-Ani [Ani70] mesurées sur des mots isolés.

L'écart entre la voyelle antérieure /a:/ et les deux autres est significatif, alors que la différence entre /u:/ et /i:/ est moins évidente (voir figure 1).

Tableau 2: Durées intrinsèques (en ms) des voyelles brèves de l'Arabe standard.

	Alg.			Syr.		
phonème	/a/	/u/	/i/	/a/	/u/	/u/
durée moy.D0	213	185	175	200	176	169
Dmin	187	142	148	180	152	146
Dmax	248	230	204	232	200	190
écart Δmax	35	45	29	32	24	23
ecart type σ	22.8	27.3	16.7	17.6	12.4	18.7

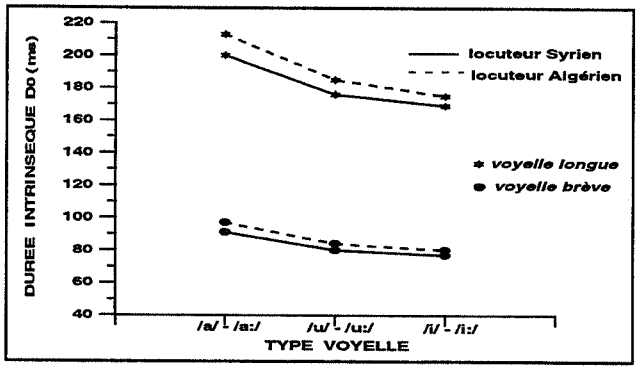


Figure 1: Durées intrinsèques des voyelles arabes

2.3 Facteur de Proportionnalité longue-brève

Les voyelles longues et leurs correspondantes brèves sont liées par un facteur de proportionnalité K, dont les valeurs sont, en moyenne, égales à 2.2 pour les locuteurs syrien et algérien.
Cette déduction doit être prise avec précaution, car les schèmes ne sont pas identiques. Aussi, une étude complémentaire portant sur les voyelles /a/ et /a:/ a été effectuée. Une liste composée de sept mots et logatomes, insérés dans une phrase porteuse /qul ...fula:n/, a été lue par trois autres locuteurs.

Tableau 3: Coefficient de proportionnalité longue/brève.

locut	Algérien1			Syrien1			Tunisien		
cont exte	/a/	/a:/	K		/a:/	K	/a/	/a:/	K
k-t	70	190	2.7	73	186	2.5	76	164	2.2
k-d	72	195	2.7	73	209	2.9	80	194	2.4
t-k	67	182	2.1	71	186	2.6	72	171	2.4
d-k	76	197	2.6	84	196	2.5	82	172	2.0
s-t	72	198	2.7	78	172	2.2	82	181	2.2
z-t	82	207	2.3	74	189	2.6	93	184	2.0
n-t	74	184	2.5	78	205	2.6	85	162	1.9

Les valeurs de K varient de 1.9 jusqu'à 2.9 et sont moins élevées chez le locuteur Tunisien. Ces valeurs sont proches de K=2, trouvée par El Ani [Ani70] avec des mots isolés et des durées nettement plus élevées. Il faut cependant relever que cette valeur de K peut prendre des valeurs extrêmes, jusqu'à K=6, pour certains types d'élocution, tels que la lecture du Coran.

Il est utile de relever que les valeurs de K, déterminées dans des conditions particulières (phrases porteuses, voyelle en position accentuée), vont subir des variations significatives avec le changement de contexte.

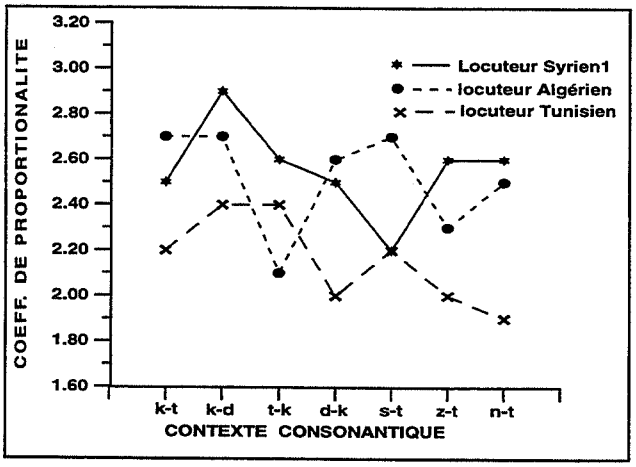


Figure 3: Coefficient de proportionnalité de durée voyelle longue-voyelle brève (opposition /a:/ - /a/).

3 Détermination du Contexte Dominant

Dans une structure C1V1C2V2C3V3, il est admis que l'effet de la consonne C2 a une influence plus importante sur la durée de la voyelle V1 que celle due à C1. Cependant, les études en Arabe sont controversées à ce sujet. Ainsi, pour Benkirane [Ben82], c'est beaucoup plus l'effet de C1 qui est plus important que celui de C2, alors que pour Mitleb [Mit84], le voisement n'a pas d'effet notable sur la durée de la voyelle, d'où une influence microprosodique négligeable. Devant des avis aussi contradictoires, nous avons d'abord tenté de déterminer le contexte dominant avec des locuteurs aux origines différentes. Les schèmes choisis sont façala, façla et

fa:çala, représentatifs, respectivement des syllabes brève ouverte, brève fermée et longue ouverte.

La détermination des contextes dominants été évaluée au moyen de deux indices: *le trait de voisement* à travers les oppositions /t/-/d/ et /s/-/z/, et *le mode articulatoire* à travers les oppositions */t/-/s/ et /d/-/z/.*

Les tableau 4 montre l'influence du contexte droit, qui a été évaluée au moyen de l'effet du trait de voisement de la consonne C2. La voyelle V1 présente un environnement C1 = /k/ et C2 est figurée par les consonnes fricatives /s/ puis /z/. L'influence du contexte gauche (tableau 5) a été évaluée en gardant C2 = /t/ et en variant la consonne C1, cette consonne étant /s/ ou /z/. Les resultats montrent que l'influence du contexte dépend du locuteur. Ainsi, le locuteur syrien1 présente un contexte droit dominant pour le schème façala, alors que pour le locuteur algérien 2, le contexte gauche est prépondérant.

L'influence du trait de voisement sur la durée de la voyelle se manifeste par un allongement temporel de cette dernière. Ceci a été observé pour le Français par Bartkhova [Bar85], pour l'Anglais par Klatt[Kla76], et pour l'Arabe par El-Ani [Ani70], ainsi que par Benkirane [ben84] pour l'Arabe Marocain. Ce phénomène a aussi été constaté dans d'autres langues au point qu'il est considéré comme un fait universel.

Tableau 4: Durée de la voyelle /a/ en fonction du contexte droit.

schème	locuteur	C2=/s/	C2=/z/	Kmi
façela	Syrien1	75	86	1.18
	Syrien2	72	84	1.17
	Algérien1	66	84	1.27
	Algérien2	66	71	1.08
façla	Syrien1	74	86	1.16
	Syrien2	77	78	1.01
	Algérien2	63	67	1.06
fa:ç ala	Syrien 1	182	191	1.05
	Algérien1	-	-	-
	Algérien2	169	171	1.01

Tableau 5: Durée de la voyelle /a/ en fonction du contexte gauche.

schème	locuteur	C1=/s/	C1=/z/	Kmi
façala	Syrien1	74	77	1.04
	Syrien2	69	84	1.18
	Algérien1	72	82	1.14
	Algérien2	67	82	1.22
façla	Syrien1	70	73	1.04
	Syrien2	65	71	1.09
	Algérien1	66	84	1.27
fa:ç ala	Syrien 1	172	190	1.10
	Algérien1	198	207	1.04
	Algérien2	164	158	0.96

4 Effet des Groupes Consonantiques

L'effet des groupes consonantiques (cluster) sur la durée de la voyelle est évalué en opposant des mots de structure C1V1C2V2C3V3 / C1V1C2C3V2 (façala/façla). Le corpus est constitué des mots suivants:

1 kataba / katba, 3 sataba / satba
2 kadaba / kadba 4 nataba / natba

Les tableaux 6a,b,c , qui expriment la variation de la durée de la voyelle /a/ en position V1, montrent que la présence d'un cluster consonantique C2C3 s'accompagne d'une diminution de la durée de la voyelle V1 qui précède celui-ci, et ce dans un rapport Kvgr. Le phénomène de la réduction de la durée précédant un cluster est beaucoup plus évident chez les locuteurs maghrébins.

Tableau 6a: Variation de la durée de la voyelle /a/ précédant un cluster (locuteur algérien)

mot	aC2V2	aC2C3	Kvgr
1	70	39	0.56
2	72	42	0.58
3	72	42	0.58
4	74	50	0.68

Tableau 6b Coefficient de réduction de la durée de la voyelle précédant un cluster (locuteur syrien)

mot	aC2V	aC2C3	Kvgr
1	73	66	0.90
2	73	70	0.96
3	74	70	0.95
4	79	77	0.97

Tableau 6b: Coefficient de réduction de la durée de la voyelle précédant un cluster (locuteur tunisien)

mot	aC2a	aC2C3	Kvgr
1	76	68	0.88
2	80	70	0.87
3	82	75	0.91
4	85	75	0.88

5 Effet des Consonnes Géminées

L'investigation précédente nous a amené à aborder l'effet des consonnes géminées et de constater que le phénomène de réduction de la durée des voyelles est également présent dans ce cas. Le sonagramme de la figure 3 montre la réduction de la durée de la voyelle V1 quand on passe du mot /badaɛa/ à /baddaɛa/. Les valeurs du tableau 7, qui regproupe les coefficients de réduction de la durée de la voyelle au voisinage de consonnes géminées, montrent aussi que la diminution de la durée de la voyelle est plus importante pour le locuteur Algérien (maghrébin), chose qui a également été constatée avec les clusters.

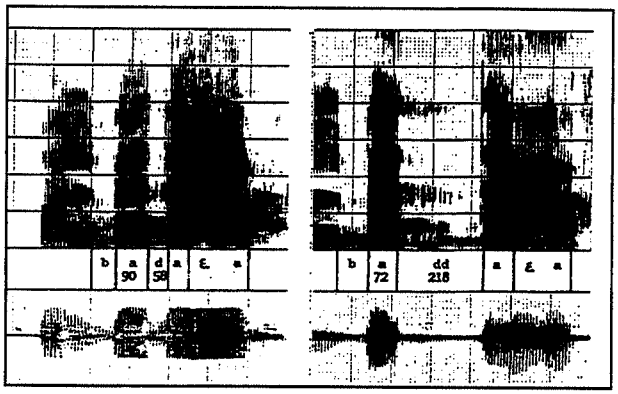


Figure 3: Réduction de la durée de la voyelle /a/ par la consonne géminée /d/.

Tableau 7: Coefficient Kvgem de réduction de V1 par une consonne géminée.

C2 géminée	Kvgem loc. Algérien	Kvgem loc.Syrien
/b/	0.46	0.85
/d/	0.49	0.88
/t/	0.51	0.83
/k/	0.50	0.91
/z/	0.60	0.83
/Z/	0.46	0.83
/f/	0.63	0.90
/ʃ/	0.62	0.87
/ħ/	0.55	0.82
/n/	0.55	0.85
/m/	0.50	0.86

Conclusion

Cette investigation dans le domaine de la prosodie de la langue arabe a permis de constater que les groupes consonantiques (cluster) et les consonnes géminées réduisent sensiblement la durée de la voyelle qui les précède.

Nous arrivons à la déduction que l'augmentation de la durée de la consonne, géminée ou première consonne du cluster, que nous avons relevé par ailleurs [Amr97] est compensée par la réduction de la durée de la voyelle précédente.

La durée de la voyelle varie aussi avec le type de consonne. L'influence d'une consonne adjacente diffère, suivant qu'elle se trouve dans une syllabe ouverte ou fermée, brève ou longue.

Les mesures des durées montrent aussi que les particularités intrinsèques des locuteurs, ou de l'origine géographique ne sont pas à négliger.

Références

Ben[82]
Benkirane M [1982], « Durée, Prosodie et Syllabation en Arabe Marocain », Publication de l'Institut de Phonétique d'Aix, Vol.8, pp45-84, 1982.

Mit[84]
Mitleb F. M. [1984], « Voicing Effect on Vowel Duration is not an Universal Absolute», Journal of Phonetics, N° 12, 1984, pp23-37.

Gha[92]
Ghazali S.[1984], «Voyelles Longues et Voyelles Brèves en Arabe Standard: Organisation Temporelle», Actes 19-èmes JEP., S.F.A., Bruxelles, 1992, pp19-24.

Mra[84]
Mrayati M., Makhoul J. [1984], « Man Machine Communication and the Arabic Language», Rapport interne, 1984.

Ani[70]
El-Ani S. [1970], « *Arabic Phonology: An Acoustical and Physiological Investigation*», Mouton the Hague, Paris, 1970.

Bar[85]
Bartkhova K. [1985], « Nouvelle Approche dans le Modèle de Prédiction Segmentale», Actes 14-èmes JEP, S.F.A., Paris, 1985.

Klat[76]
Klatt D.H., [1976], «Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence», J.A.S.A., Vol. 59, N°5, 1976, pp1208-1221.

Amr[97]
Amrouche A., Boudraa B., Debyeche M., Rouvaen J.M.,[1997], «Prédiction des Durées des Phonèmes en vue de la Synthèse de la Parole en Arabe Standard», Actes 17-èmes J.T.E.A., Nabeul, 1997, pp318-325.

Prosodie

Analyse multi-paramétrique des configurations tonales des proéminences en français

C. Astesano , R. Espesser, E. Flachaire, P. Nicolas

Institut de Phonétique, ESA 6057 "Parole et Langage"
Université de Provence, 29, Avenue Robert Schuman
13621 Aix-en-Provence
Fax: + 33 04 42 59 50 96
e-mail : corine.astesano@lpl.univ-aix.fr

ABSTRACT

This paper presents a description of tonal configurations of prominence in French based on a multi-style corpus. This study is carried out within the typological framework developed in Aix depicting prominence on a functional basis. The phonetic correlates of the pre-defined accentual categories are investigated according to the dynamic properties of tonal variations, namely the slopes, timing and span of the f0 contours. This phonetic description is complemented by an attempt to predict membership of prominence for a particular accentual category using a regression model. This empirical investigation clearly validates the functional distinction of prominences on a tonal basis, although the inclusion of duration parameters would unquestionably improve the global predictions of the model.

1. INTRODUCTION

Diverses études s'attachent à décrire le système accentuel du français ([Fon80],[Ros80],[Fle91],[Vih93],[Dic97]) sans toutefois parvenir à une description fédératrice et consensuelle. De plus, les tentatives d'applications des modèles descriptifs à la parole spontanée ont mis en évidence la complexité de l'organisation accentuelle du français et ont révélé la nécessité de modèles adaptatifs. Lors d'une étude précédente [Ast95], nous avions validé empiriquement l'existence des catégories accentuelles du français telles qu'elles sont définies dans le cadre typologique développé à Aix [Dic97],[Dic(forth)]. Cette description postule l'existence d'un système accentuel complexe en français, fondé sur deux classes accentuelles principales : la classe rythmique, qui comprend l'accent secondaire initial et l'accent final de syntagme, et la classe sémantico-pragmatique qui comprend l'accent emphatique et l'accent nucléaire associé à une frontière prosodique terminale ou non terminale.

Nous proposons ici de poursuivre l'étude des corrélats phonétiques de ces catégories formelles et fonctionnelles pré-définies, en terme de propriétés tonales. En effet, si le français est généralement défini sur une base principalement temporelle, des études récentes ([Dic97], [Bec93]) ont mis l'accent sur le rôle des indices tonals dans l'organisation rythmique du français, dans la mesure où certaines syllabes dans un mot peuvent être associées phonologiquement à un ton ou une séquence de tons. Terken [Ter93] insiste sur le rôle prépondérant de la hauteur dans la perception de la proéminence. Il est cependant partisan de l'étude combinée des facteurs prosodiques et notamment des facteurs liés à l'aspect dynamique des variations de f0. Dans le même ordre d'idée, t'Hart & al. [Har90] fondent leur description en contours tonals (plutôt qu'en tons discrets) sur le caractère continu et dynamique de la f0.

Nous choisissons de prendre en compte les propriétés temporelles de la f0 telles que les directions de pentes, l'empan syllabique et le timing des configurations, afin de rendre compte des spécificités de l'organisation tonale du français. Cette analyse est complémentaire aux études précédentes de l'organisation tonale du français en terme de tons discrets [Hir(forth)]. Notre but est donc double : nous proposons d'une part un examen des caractéristiques phonétiques de catégories accentuelles pré-définies en terme de configurations tonales locales ; nous voulons déterminer d'autre part dans quelle mesure la combinaison des différentes propriétés tonales retenues permet de prédire l'appartenance d'une proéminence particulière à une catégorie accentuelle donnée.

2. MATERIEL ET PROCEDURE

2.1. Matériel

Notre étude a été réalisée sur 10 minutes de parole connectée, constituée de trois styles de parole (lecture, journal radiophonique, interview) comptant chacun 2 locuteurs. Le choix de 3 styles de parole différents nous permet de mettre en valeur les invariants dans les réalisations prosodiques des proéminences. Les enregistrements ont été transcrits sans ponctuation. Un expert a localisé les proéminences perceptivement, indiqué les accents emphatiques et les frontières d'unités intonatives terminales et non terminales. Le corpus comprend un total de 786 proéminences perçues, classées parmi les catégories accentuelles suivantes : accent emphatique initial (EMP), accent final d'unité intonative non terminale (UI-N), accent final d'unité intonative

terminale (UI-T), accent initial de mot (AI) et accent final de mot (prosodique) (AF). Il faut noter que les occurrences de proéminences ne sont pas distribuées de manière homogène dans les catégories accentuelles, les AF étant largement plus représentés que les UI-T ou les EMP par exemple.

2.2. *Procédure expérimentale*

Les extraits ont été numérisés à 16 kHz et un étiquetage phonémique et syllabique manuel a été réalisé. La f0 est modélisée par une séquence de points cibles <ms ;Hz> correspondant aux variations locales pertinentes [Hir93]. Les configurations tonales sont définies par une série de 3 points cibles. Le point cible le plus proche de l'attaque vocalique de la syllabe proéminente constitue le centre de la configuration tonale, dont les frontières sont délimitées par les points cibles gauche et droit adjacents au point cible central. Les configurations tonales dont les frontières droites ou gauches s'étendaient au delà d'une pause silencieuse ont été éliminées. Sur la base de ces configurations, les 8 paramètres suivants ont été extraits (figure 1) : l'empan syllabique des configurations tonales, les distances entre les 3 points cibles en terme de ms et de valeurs ERB [Her91] (en valeurs absolues), les pentes droites et gauches (ERB/ms), et la distance du point cible central à l'attaque vocalique de la syllabe proéminente.

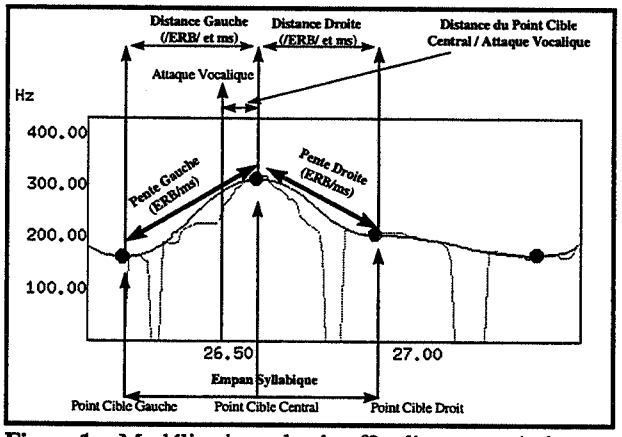


Figure1: Modélisation de la f0 d'une proéminence perçue. 3 points cibles définissent la configuration tonale dont sont extraits les 8 paramètres.

2.3.*Méthodologie statistique*

Notre but est de savoir si les mesures (8 paramètres tonals) permettent la distinction des catégories accentuelles. Le traitement statistique d'une analyse multi-paramétrique peut se faire d'un point de vue descriptif à l'aide d'analyses factorielles (Analyse Discriminante ou Analyse en Composantes Principales), ou d'un point de vue quantitatif avec une Analyse de Régression des variables dépendantes qualitatives. Le *modèle logit multinomial* permet d'obtenir des probabilités d'appartenance à un groupe (ici, catégories accentuelles) en fonction des mesures explicatives. Cette méthode est toutefois coûteuse lorsque le nombre de variables est important. Afin de réduire le nombre de variables étudiées, la sélection des paramètres tonals à intégrer dans le *modèle logit* s'est faite à l'aide d'une *Analyse de Variance*. L'Anova est également à la base de la description phonétique des configurations tonales. Un pré-traitement des variables a été nécessaire avant l'étape d'analyse : les valeurs en Hz ont été transformées en ERB, nous permettant ainsi d'interpréter les variations de f0 sur une échelle perceptive. De plus, la transformée en Z des valeurs de ERB nous a permis de comparer les variables indépendamment du locuteur ou du style de parole (fluctuation à long terme de la f0 et registre du locuteur). Les variables pertinentes à retenir pour le modèle *logit* sont donc les paramètres indépendants du locuteur permettant de distinguer les catégories accentuelles. L'interaction entre les deux variables indépendantes (locuteurs et catégories accentuelles) devait de surcroît être non significative pour augmenter la fiabilité des résultats (table 1).

Table 1 : Paramètres pertinents pour la description des proéminences, à inclure dans le modèle logit. (Cat.Acc = catégories accentuelles ; Loc. = locuteurs ; Interac. = interaction entre catégories accentuelles et locuteurs).

	Pente Gauche (ERB/ms)		Pente Droite (ERB/ms)		Ecart Gauche ERB		Empan Syllabique	
	P-val	F-val	P-val	F-val	P-val	F-val	P-val	F-val
Cat.Acc	<,0001	16,67	,0050	4,329	<,0001	11,75	,0175	3,405
Loc.	,5529	,796	,4547	,940	,1615	1,589	,5149	,850
Interac	,0553	1,660	,0432	1,723	,0937	1,519	,8578	,622

Dans le modèle logit, la quantification de l'effet des variables explicatives est donnée par un *pseudo-R^2* : une valeur proche de 0 signifie que les variables explicatives ne permettent pas de distinguer les catégories, alors qu'elles sont clairement distinguées si la statistique est proche de 1. Nous avons choisi le R^2 de Nagelkerke [Nag91] qui pallie la minoration systématique des R^2 de type Mac Fadden [Fad74].

3 RESULTATS

3.1.*Analyse de Variance*

La catégorie UI-N avait un comportement instable d'un locuteur à l'autre, et a donc été éliminée. On pourrait rapprocher cela des observations de Vaissière [Vai83] selon laquelle le rôle de la f0 et de la durée varie en fonction du locuteur dans les tâches de segmentation de la parole en unités prosodiques. La fonction sémantico-pragmatique de la catégorie UI-N la rend sans doute particulièrement sensible aux stratégies individuelles de marquage prosodique. Au vu des résultats de l'anova, nous conservons dans le modèle logit les 4 paramètres

suivants : pentes droite et gauche, écart gauche |ERB| et empan syllabique (table 1). Il est intéressant de noter que les paramètres comprenant uniquement des durées brutes (timing) se sont avérés non pertinents. Apparemment, il existerait une 'normalisation' inter-locuteurs en terme de syllabes (empan syllabique) mais pas en terme de durée brute, ce qui semble confirmer le rôle de la syllabe comme unité de planification.

3.2.Modèle Logit Multinomial

Nous prenons en compte dans le modèle 4 groupes d'accents distincts [*AF* (G_0), *UI-T* (G_1), *AI* (G_2) and *EMP* (G_3)] et 4 variables explicatives [*pente gauche* (x_1), *pente droite* (x_2), *écart gauche* |*ERB*| (x_3) et *empan syllabique* (x_4)]. Le corpus est maintenant constitué de 543 données, $372 \in G_0$, $36 \in G_1$, $81 \in G_2$, $53 \in G_3$.

Table 2 : Résultats des pseudo-R^2

R^2_{all}	R^2_{G0G1}	R^2_{G0G2}	R^2_{G0G3}	R^2_{G1G2}	R^2_{G1G3}	R^2_{G2G3}
0.56	0.32	0.13	0.17	0.52	0.73	0.22

La table 2 présente les résultats du pseudo-R^2 du modèle multinomial (R^2_{all}) et ceux des modèles binomiaux (catégories accentuelles deux à deux). La catégorie UI-T (*G1*), étant donné son contour globalement descendant, se distingue clairement dans les oppositions de catégories deux à deux (modèles binomiaux). Le modèle général explique 56% de la variabilité ; les paramètres tonals semblent donc avoir un effet explicatif non négligeable dans la prédiction d'appartenance d'une proéminence à une catégorie accentuelle.

4 DISCUSSION

Individuellement, les 4 paramètres permettent de distinguer clairement les catégories accentuelles (tests post-hocs de l'anova). Pour le modèle logit général, les statistiques de Student montrent pour la catégorie UI-T notamment, que le rejet de l'hypothèse de non-influence est le mieux testé par la pente gauche (t_{x1}=-6,2). En combinaison dans les modèles logits binomiaux néanmoins, les paramètres pertinents diffèrent suivant les catégories opposées.

Considérant seulement les comparaisons phonologiques pertinentes, deux résultats majeurs émergent de ces analyses : la confirmation de la fonction rythmique des proéminences initiales et finales de mots (AI et AF), ainsi que la distinction - relative - entre les proéminences initiales rythmiques et emphatiques (AI et EMP).

Fonction rythmique des AI et AF (G_0G_2) : Di Cristo & Hirst [Dic97] font état de la 'visée eurythmique' des proéminences initiales de mots et finales de mots (prosodiques). La tendance à éviter la contiguïté accentuelle ou à favoriser l'équilibre rythmique inter-accentuel en terme de syllabes, s'actualise à un niveau de surface par l'occurrence alternée de proéminences initiales et finales de mots à l'intérieur de l'Unité Intonative. Les propriétés rythmiques de AI et AF se manifestent de plus au niveau tonal par une tendance à former un 'arc accentuel' [Fon80] (pour une discussion, voir [Dic(forth)]). Nos résultats confirment ces caractéristiques à la fois du point de vue de l'empan syllabique et des configurations tonales générales de ces catégories accentuelles. Ces deux types de proéminences ont en effet les empans syllabiques les plus étendus, ce qui signifie qu'aucune variation de f0 majeure ne se produit dans leur voisinage. Cette tendance confirme donc l'hypothèse d'eurythmie. La pente gauche ne distingue pas AF et AI (Tests post-hoc Anova : p-value=0.4722 ; Logit Binaire : T-student=1.38). En revanche, la pente droite les distingue significativement (p-value=0.001 ; T-student=3.91). La configuration tonale de AF est symétrique, alors que celle de AI est asymétrique (pente droite en moyenne deux fois moins abrupte que pente gauche). Cette particularité de configuration asymétrique pourrait refléter le rôle de AI dans la démarcation à gauche des unités lexicales, ainsi que la réalité phonétique de l'arc accentuel. Ces résultats nous amènent également à une comparaison avec des études antérieures [Jun95]. Dans une certaine mesure, et malgré nos approches méthodologiques différentes, ils semblent corroborer la description de ce que ces auteurs appellent 'le syntagme accentuel' du français (Accentual Phrase). La question reste de savoir si ces similarités confirment ou non l'existence, à un niveau phonologique profond, de ce syntagme accentuel borné à droite et à gauche (pour une discussion plus détaillée, voir [Hir96]). Enfin, le modèle logit binaire opposant ces deux catégories accentuelles présente un R^2 faible (R^2_{G0G2} = 0.13), ce qui signifie que, même si la pente droite est un paramètre pertinent dans leur distinction, la *combinaison* des 4 paramètres tonals ne permet pas de les distinguer nettement; leur similitude fonctionnelle se traduit en effet au plan phonétique par un comportement tonal quasi identique.

Proéminences initiales rythmiques (AI) et emphatiques (EMP) (G_2G_3): Les 4 paramètres tonals distinguent significativement les deux types de proéminences (tests post-hocs de l'anova), confirmant par la même leurs rôles distincts sur le plan fonctionnel (voir [Ast95], [Dic(forth)] pour une discussion sur leur assimilation fréquente dans la littérature). En effet, la configuration tonale de AI n'est pas très saillante et s'étend sur un grand nombre de syllabes, alors que celle de EMP est très saillante et a un empan syllabique réduit. Elles ont toutes deux la même configuration générale (asymétrique), celle de EMP étant en moyenne 2 fois plus grande que celle de AI. Le modèle logit binaire révèle cependant que seuls l'empan syllabique (T-student=-2.82) et la pente droite (T-student=-2.34) sont pertinents dans la distinction des catégories. Les paramètres expliquent une part non négligeable de la variabilité du modèle (R^2_{G2G3} = 0.22) ; cette valeur néanmoins assez faible est sans doute due à la similitude dans l'aspect

général de leurs configurations tonales. Di Cristo [Dic(forth)] décrit d'ailleurs l'accent emphatique initial comme étant l'expression 'hyper' de l'accent initial rythmique, hypothèse confirmée par nos résultats. Il apparaît donc qu'un changement graduel dans les corrélats phonétiques induit un changement dans la catégorisation fonctionnelle de ces proéminences.

Table 3 : Résumé des caractéristiques tonales des catégories accentuelles.

Catégories Accentuelles	Distance Gauche ERB	Configurations Saillantes	Symétrie des Configurations	Empan Syllabique
AF	+	-	symétrique	+
UI-T	- -	- -	symétrique	-
AI	+ +	-	asymétrique	+ +
EMP	+ + +	+ +	asymétrique	- -

5. CONCLUSION

La combinaison des deux approches statistiques (analyse de variance et modèle de régression) nous a permis d'avoir une vision plus précise et plus complète des spécificités des configurations tonales du français. En effet, le modèle de régression logit, en prenant en compte la *combinaison* des paramètres tonals et en *quantifiant* leurs rôles relatifs dans la distinction des catégories accentuelles, nous a permis de rendre compte précisément des similitudes et différences phonétiques, corollaires des rôles fonctionnels des catégories accentuelles. La capacité du modèle logit général à expliquer 56% de la variabilité sur la seule base des paramètres tonals nous invite à poursuivre dans cette approche. Une étude sur l'intégration de paramètres temporels (durée syllabique) est actuellement en cours [Ast(forth)].

BIBLIOGRAPHIE

[Fon80] Fónagy, I., 1980, "L'accent en français : accent probabilitaire", Studia Phonetica, Vol.15, pp. 123-233.

[Ros80] Rossi, M., 1980, "Le français, langue sans accent?", Studia Phonetica, Vol. 15, pp. 15-51.

[Fle91] Fletcher, J., 1991, "Rhythm and final lengthening in French", Journal of Phonetics, Vol.19.2, pp.193-212

[Vih93] Vihanta, V., 1993; "Focalisations et autres proéminences en français lu et spontané", Mélanges Lingren, Turun

[Dic97] Di Cristo, A. & Hirst, D. J., 1997, "L'accent en français : stratégies et paramètres", in J. Perrot (ed.) Polyphonie pour Ivan Fónagy

[Ast95] Astésano, C., Di Cristo, A., & Hirst, D., 1995, "Discourse-based empirical evidence for a multi-class accent system in French", Proc. of 13th ICPhS, Vol. 4, pp. 630-633, Stockholm

[Dic(forth)] Di Cristo, A., (forthcoming);"Vers une modélisation de l'accentuation en français", T.I.P.A n°17

[Bec93] Beckman, M., 1993, "Modeling the Production of Prosody", Working Papers, ESCA, n°41, pp. 258-263, Lund

[Ter93] Terken, J., 1993, "Issues in the Perception of Prosody", Working Papers, ESCA, n°41, pp. 228-233, Lund

[Har90] t'Hart, J., Collier, R., & Cohen, A., 1990, A perceptual study of intonation, CUP.

[Hir(forth)] Hirst, D., Di Cristo, A. & Espesser, R. (forthcoming) : "Levels of representation and levels of analysis for the description of intonation systems", in Horne, M. (ed.), Prosody : Theory & Experiment, Dordrecht : Kluwer Academic Publishers

[Hir93] Hirst, D. J. & Espesser, R., 1993, "Automatic modelling of fundamental frequency with a quadratic spline function", T.I.P.A, Vol. 15, pp. 71-85

[Her91] Hermes, D.J. & Van Gestel, J.C., 1991, "The frequency scale of speech intonation", J.Acoust. Soc. Am., 90 (1), pp. 97-102

[Fad74] McFadden , 1974a, "Conditional logit analysis of qualitative choice behaviour", in Zarembka, P.(ed.), Frontiers in Econometrics, N.Y., Academic Press

[Nag91] Nagelkerke, N.J.D., 1991 "A note on a general definition of the coefficient of determination", Biometrica, Vol. 78, n° 3, 691-692

[Vai83] Vaissière, J., 1983, "Language-Independent Prosodic Features", in Cutler, A. & Ladd, D.R. (eds), 1983, Prosody: Models and Measurements, Heidelberg, pp.53-66

[Jun95] Jun, S. & Fougeron, C., 1995, "The Accentual Phrase and the prosodic structure of French", Proc. of 13th ICPhS, vol. 2, pp. 722-725, Stockholm

[Hir96] Hirst, D., Di Cristo, A., 1996, "Y a-t-il des unités tonales en français? ", XXI^{es} J.E.P., Avignon, 10-14 Juin 1996, pp. 223-226

[Ast(forth)] Astésano, C., (en cours) "Approche paramétrique de l'organisation rythmique du français contemporain", Thèse en cours, Université de Provence, Aix.

Evaluation de modèles d'étiquetage automatique de l'intonation

Estelle Campione, Emmanuel Flachaire, Daniel Hirst, Jean Véronis

Laboratoire Parole et Langage,
Université de Provence & CNRS
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
Tel. : +33 4 42 95 36 33, Fax : +33 4 42 59 50 96
E-mail: Estelle.Campione@lpl.univ-aix.fr

ABSTRACT

In this paper we discuss various problems related to the automatic labelling of F0 contours through the evaluation of five models of increasing complexity. The models were tested on a French corpus and the best two models were subsequently evaluated for Italian (twenty speakers for a total of 90 minutes of speech). Although there is still some room for improvement, the best two models ensure a quite satisfactory resynthesis of the original F0 contours (mean quadratic error of about one semi-tone on either absolute F0 values or relative pitch intervals).

1. INTRODUCTION

L'étiquetage prosodique de grands corpus serait extrêmement utile pour l'étude de la parole en général, ainsi que pour des tâches d'ingénierie telles que la génération d'une prosodie de qualité en synthèse [Ver97] ou bien de l'utilisation de la prosodie en reconnaissance. Les quelques corpus qui existent (par exemple [Ost95]) ont été étiquetés manuellement par des experts. Outre le caractère lent et difficile de cette tâche, son caractère éminemment subjectif diminue la fiabilité des résultats, ou impose le recours à des contre-expertises qui augmentent les coûts. Il serait donc extrêmement intéressant de disposer de systèmes d'étiquetage automatique permettant de s'affranchir de la phase d'intervention manuelle, ou de la réduire à une simple phase de vérification et de correction. Toutefois, on ne dispose pas, à l'heure actuelle, d'étiqueteur prosodique automatique fiable, même pour les systèmes de transcription les plus utilisés, tels que ToBI, bien que des recherches soient en cours ([Bla96], [Dus97]).

Nous discutons dans cette communication de divers problèmes liés à l'étiquetage automatique de l'intonation, à travers la présentation de cinq modèles de complexité croissante. Les cinq modèles ont été testés sur un corpus français, et les deux meilleurs modèles ont été ensuite été évalués sur l'italien (20 locuteurs et 90 minutes de parole au total). Bien que perfectibles sur divers aspects, les deux meilleurs modèles proposés permettent une bonne régénération des contours intonatifs originaux (erreur quadratique moyenne inférieure au demi-ton sur les hauteurs absolues ou les intervalles mélodiques).

2. PROBLEMATIQUE

On considère souvent que la F_0 est la combinaison d'une composante macroprosodique qui reflète le choix d'un patron intonatif par le locuteur, et d'une composante microprosodique qui ne résulte pas d'un choix du locuteur (abaissement des constrictives, etc.) [DiC86]. De nombreuses études se sont attachées, depuis les années 60, à factoriser ces deux composantes, et à extraire le plus automatiquement possible l'information macroprosodique pertinente du signal de parole. Cette extraction peut être décomposée en deux étapes :

- *stylisation,* c'est-à-dire remplacement de la courbe de F_0 par une fonction numérique plus simple conservant la même information macroprosodique ;
- *codage symbolique,* c'est-à-dire représentation par une suite de symboles constituant une discrétisation de la courbe stylisée.

La première étape est parfois qualifiée de *close-copy stylisation* ([DeP79]), dans la mesure où il est souhaitable que des énoncés re-synthétisés, dans lequel la F_0 serait remplacée par la F_0 stylisée, ne soient pas perceptuellement distincts de l'original. Cette notion de *close copy* peut être étendue au codage symbolique : il serait souhaitable que la discrétisation effectuée ne perde pas d'information pertinente en ce qui concerne la macroprosodie, et que la F_0 régénérée à partir du codage symbolique (à l'aide d'un modèle adéquat) ne se distingue pas non plus perceptuellement de la courbe stylisée et de l'original. On aurait ainsi un système de stylisation et de codage totalement réversible, qui serait du plus grand intérêt pour le codage automatique de grands corpus de parole.

La stylisation a fait l'objet de nombreux travaux (par exemple : [tHa90], [Hir93], [Tay94], [DAl95], etc.), et on peut dire que la technique en est maîtrisée de façon satisfaisante. Divers systèmes de codage symbolique ont également été proposés, dont le plus connu est sans doute ToBI [Sil92]. Toutefois, l'automatisation et la réversibilité (*close copy*) sont beaucoup moins avancées pour le codage symbolique que pour la stylisation.

3. ASPECTS COMMUNS AUX 5 MODELES

Les cinq modèles testés utilisent tous la même méthode de stylisation automatique, MOMEL (MOdélisation MELodique), a été proposée par Hirst et Espesser [Hir93] (voir aussi [Hir98a]). Contrairement à la plupart des autres méthodes de stylisation, MOMEL utilise des fonctions splines quadratiques (c'est-à-dire des arcs de parabole), lisses en tout point, et non des segments des droites, qui introduisent des angles vifs consécutifs. De plus, les segments non voisés sont interpolés et la courbe stylisée ne présente donc aucune discontinuité. Les courbes splines quadratiques utilisées peuvent être représentées par une suite de *points-cibles* correspondant aux seuls changements significatifs (passages par zéro de la tangente), à partir desquels le codage symbolique peut être calculé.

Le plus connu des systèmes de codage symbolique est ToBI [Sil92], utilisé avec succès par de nombreux auteurs sur l'anglais. Toutefois, ToBI présente l'inconvénient d'être un système spécifique à une langue : son adaptation à d'autres langues, bien que possible, nécessite un effort assez important. Par ailleurs, l'étiquetage ToBI repose sur des jugements linguistiques pris par un expert, et est difficilement automatisable (bien que des tentatives existent dans ce sens : [Bla96]). Enfin, la régénération d'une F_0 proche de l'original à partir de ToBI est délicate.

Les modèles présentés dans cette étude sont inspirés par le système de codage symbolique INTSINT (INternational Transcription System for INTonation) proposé par Hirst et Di Cristo [Hir98b]. Au contraire de ToBI, qui encode des événements de nature linguistique, INTSINT est un codage purement formel de la courbe macroprosodique (cette idée est aussi utilisée par Taylor [Tay93], dans le système HLCB). Chaque symbole permet de coder un point cible, soit de façon absolue, soit par rapport au registre du locuteur ou bien de façon relative par rapport aux points environnants. Ce type de codage formel peut être considéré comme un premier niveau de codage, pouvant servir de base, le cas échéant, à des systèmes de plus haut niveau tels que ToBI.

Comme dans le système INTSINT, tous les modèles testés comportent deux symboles absolus *T* (*top*) et *B* (*bottom*) qui servent à coder les points extrêmes, ainsi qu'un symbole relatif *S* (*same*) indiquant une absence de mouvement mélodique :

- les points cibles dépassant un seuil τ_T sont codés *T*, ceux au-dessous d'un seuil τ_B sont codés *B* (τ_T et τ_B sont choisis de façon que 5% des points soient codés *T* et autant soient codés *B* en supposant une distribution normale) ;
- les points cibles situés à moins de 2.5% (en demi-tons ou ST) du point précédent sont codés *S*.

Les modèles différent par le codage relatif des mouvements ascendants et descendants (voir sous-section suivante). Le premier point, s'il n'est pas codé *T* ou *B* est codé par rapport à la moyenne des fréquences des points-cibles du segment de parole modélisé.

La régénération s'effectue à partir d'un point fictif situé avant le début de l'énoncé et de fréquence égale à la moyenne. Les points suivants sont générés de la façon suivante :

- les points *T* correspondent à une fréquence F_T et les points *B* à une fréquence F_B respectivement déterminées par la moyenne des points au-dessus de τ_T et au-dessous de τ_B dans une distribution normale ;
- les points *S* gardent la fréquence du point précédent ;
- les fréquences des autres points relatifs sont calculés par régression linéaire à partir du point précédent, à l'aide de coefficients de régression estimés pour chaque symbole sur le corpus à modéliser (tous locuteurs confondus en valeurs centrées réduites pour une langue donnée).

4. SPECIFICITES

4.1. Modèle HL

Le premier modèle testé est un modèle minimal qui comporte seulement deux symboles en plus des deux symboles *T*, *B* et *S* communs à tous les modèles :

- *H* : montant ;
- *L* : descendant.

4.2. Modèle Config

Le second modèle testé est un modèle dont les symboles utilisés représentent la configuration locale des points cibles:

- *H* : pic ;
- *U* : plateau montant (*upstepped*) ;
- *D* : plateau descendant (*downstepped*) ;
- *L* : vallée.

4.3. Modèle Ampli2

Le modèle *Ampli2* est basé sur des symboles représentant les amplitudes des mouvements ascendants et descendants :

- H_2 : ascendant d'amplitude supérieure à un seuil α ;
- H_1 : ascendant d'amplitude inférieure à α ;
- L_1 : descendant d'amplitude inférieure à α ;
- L_2 : ascendant d'amplitude supérieure à α .

Par essais systématiques, nous avons déterminé que les meilleurs résultats de ce modèle sont obtenus lorsque α correspond à 40% de $\tau_T - \tau_B$.

4.4. Modèle Ampli3

Le modèle *Ampli3* est basée sur la même idée mais utilise trois niveaux d'amplitude déterminés par deux seuils α_1 et α_2 :

- H_3 : ascendant d'amplitude supérieure à α_2;
- H_2 : ascendant d'amplitude entre α_1 et α_2 ;
- H_1 : ascendant d'amplitude inférieure à α_1 ;
- L_1 : descendant d'amplitude inférieure à α_1 ;
- L_2 : descendant d'amplitude entre α_1 et α_2 ;
- L_3 : descendant d'amplitude supérieure à α_2.

Les meilleurs résultats de ce modèle sont obtenus lorsque α_1 et α_2 correspondent respectivement à 25% et 50% de $\tau_T - \tau_B$.

4.5. Modèle Niveaux

Le modèle *Niveaux* utilise un principe différent. La bande centrale entre les deux seuils τ_T et τ_B est divisée en trois bandes, correspondant chacune à 30 % des points dans une distribution normale : *G* (Grave), *M* (Médium), *A* (Aigu). Le codage des points relatifs se fait en associant la direction par rapport au point précédent et la bande dans laquelle se trouve le point :

- H_A : ascendant, aboutissant dans l'aigu ;
- H_M : ascendant, aboutissant dans le médium ;
- H_G : ascendant, aboutissant dans le grave ;
- L_A : descendant, aboutissant dans l'aigu ;
- L_M : ascendant, aboutissant dans le médium ;
- L_G : ascendant, aboutissant dans le grave.

5. EVALUATION

5.1. Corpus

Les cinq modèles ont d'abord été évalués sur le français, puis les deux meilleurs modèles ont ensuite été testés sur l'italien. Le corpus utilisé est une partie de la base de données EUROM1 [Cha95], comprenant 35 minutes de parole pour le français et 54 pour l'italien. Les phrases de ce corpus sont regroupées en passages de 5 phrases prononcés par 10 locuteurs par langue (5 hommes et 5 femmes).

La totalité du corpus a été stylisé automatiquement a l'aide de l'algorithme MOMEL, puis vérifié et corrigé manuellement par des experts. Le corpus contient 6329 points cibles pour le français et 9804 pour l'italien. Les fréquences de points cibles ont été converties en demi tons (ST). L'écart type varie de 2.79 ST à 4.18 ST pour les locuteurs français et de 2.78 ST à 4.44 ST pour les locuteurs italiens. Les points initiaux des énoncés d'un locuteur ont une moyenne très proche de la moyenne de l'ensemble des points cibles de ce locuteur (écart moyen de –0.88 ST en français et 0.34 ST en italien). La distribution peut être approximée de façon grossière par une loi normale.

5.2. Méthode

Les courbes régénérées ont été comparées avec les courbes stylisées vérifiées par des experts. Comme il a été souvent noté dans la littérature, la mesure de la similarité entre deux courbes intonatives n'est pas simple, et en tout état de cause devrait faire intervenir des tests perceptifs, rarement possibles à grande échelle. Nous utiliserons comme mesures :

- e_{abs} : erreur quadratique moyenne sur les fréquences des points-cibles (en ST) ;
- e_{rel} : erreur quadratique moyenne sur l'amplitude des mouvements mélodiques (en ST) ;
- p_{abs} : pourcentage de points-cibles à moins de 2 ST de la fréquence originale ;
- p_{rel} : pourcentage de mouvements mélodiques à moins de 2 ST de l'amplitude originale.

5.3. Résultats

Les résultats sur le français sont récapitulés par les figures 1 et 2. La valeur de e_{abs} est minimale pour le modèle *Niveaux* (0,96 ST), celle de e_{rel} est minimale pour le modèle *Ampli3* (1,09 ST). Le pourcentage p_{abs} est maximal pour le modèle *Niveaux* (96,0%), et le pourcentage p_{rel} est maximal pour le modèle *Ampli3* (93,1%).

Les résultats obtenus sur l'italien en ce qui concerne les deux meilleurs modèles, seuls testés sur cette langue, sont analogues.

6. DISCUSSION

Comme il est prévisible, les résultats s'améliorent avec le nombre de symboles utilisés. On note cependant les points suivants :

- Le modèle *HL* est clairement trop fruste pour être utilisable de façon pratique.
- A nombre de symboles égaux, le modèle *Ampli2* surpasse le modèle *Config*. Il ne semble donc pas que la configuration (pic, plateau) soit porteuse de suffisamment d'information pour une bonne régénération. Au demeurant, la configuration peut être dérivée de façon triviale du modèle *Ampli2* par l'examen de la séquence de symboles.
- Enfin, les deux meilleurs modèles, *Ampli3* et *Niveaux*, de nombre de symboles égaux, ont un comportement opposé : *Ampli3* permet une bonne restitution des mouvements mélodiques, mais une moins bonne restitution des hauteurs absolues, alors que *Niveaux* se comporte de façon inverse. On note également que les résultats sont similaires

en français et en italien, ce qui laisse penser que les modèles concernés sont relativement indépendants de la langue.

Une discrimination entre les deux meilleurs modèles demanderait des tests perceptifs de façon à déterminer s'il est préférable d'optimiser la restitution des hauteurs absolues ou celle des intervalles. Ces modèles sont de plus susceptibles d'améliorations diverses. Il semblerait en particulier que des symboles additionnels soient nécessaires pour coder les points les plus extrêmes du registre du locuteur, qui sont les points les plus mal régénérés dans tous les modèles.

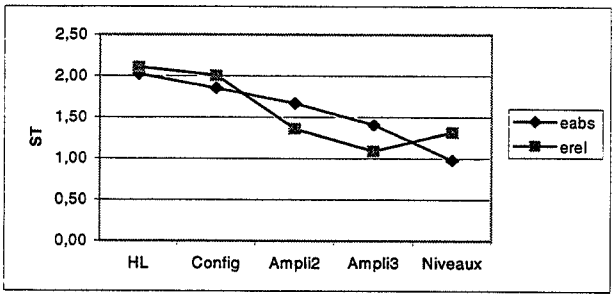


Figure 1: Erreur quadratique moyenne (français)

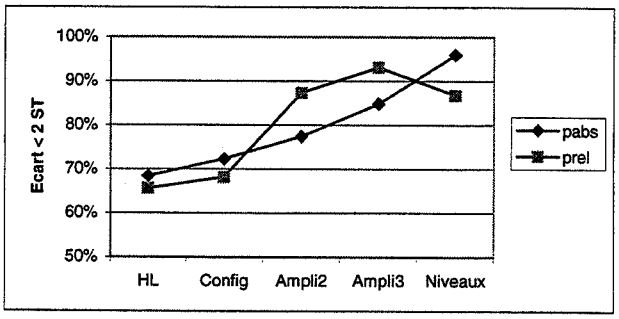


Figure 2: Pourcentage d'écarts <2ST (français)

REMERCIEMENTS

Les auteurs remercient Robert Espesser pour son aide technique, Corine Astésano et Fabienne Courtois pour leur travail de correction du corpus.

BIBLIOGRAPHIE

[Bla96] Black, A., Hunt, A. (1996). Generating F_0 contours from ToBI labels using a linear regression, *Proc. ICSLP'96*, Philadelphia.

[Cha95] Chan, D., *et al.* (1995). EUROM- A Spoken Language Resource for the EU. *Proceedings of Eurospeech'95.* Madrid, 1, 867-870.

[DAl95] D'Alessandro, C., Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language*, 9, 257-288.

[DeP79] De Pijper, JR. (1979). Close-copy stylisation of British English intonation contour, *IPO Annual Progress Report* 14, 66-71.

[DiC86] Di Cristo, A. Hirst, D.J. (1986). Modelling French micromelody : analysis and synthesis. *Phonetica,* 43, 1/3, 11-30.

[Dus97] Dusterhoff, K., Black, A. (1997). Generating F_0 Contours for Speech Synthesis Using the Tilt Intonation Theory, *Proceedings of an ESCA Tutorial and Research Workshop on Intonation : Theory, Models and Applications*, September 1997, Athens, 107-110.

[Hir93] Hirst, D., Espesser, R. (1993). Automatic Modelling of Fundamental Frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix-en-Provence,* 15, 75-85.

[Hir98a] Hirst, D., Di Cristo, A., Espesser, R. (forthcoming). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (Ed.), *Prosody: Theory and Experiment,* Dordrecht: Kluwer Academic Publishers.

[Hir98b] Hirst, D., Di Cristo, A. (in press) A survey of intonation systems. in Hirst, D., Di Cristo, A. (Eds) *Intonation Systems: a Survey of Twenty Languages.* Cambridge : Cambridge University Press, 1-44.

[Ost95] Ostendor M.F., Price P. J., Shattuck-Hufnagel S. (1995). The Boston University Radio News Corpus, , *Technical Report No. ECS-95-001,* Boston University.

[Ros72] Rossi, M., Chafcouloff, M. (1972). Les niveaux intonatifs. *Travaux de l'Institut de Phonétique d'Aix,* 1,167-176.

[Sil92] Silverman, K., *et al.* (1992). ToBI: a standard for labelling English prosody. *Proc. ICSLP'92,* 2, 867-870, Banff, Canada.

[Tay93] Taylor, P. (1993). Automatic Recognition of Intonation from F0-contours using the Rise/Fall /Connection Model, *Proceedings of Eurospeech'93*, Berlin, 2, 789-792.

[Tay94] Taylor, P. (1994). The Rise/Fall/ Connection Model of Intonation, *Speech Communication,* 15:1&2, 169-186.

[tHa90] t'Hart, J., Collier, R., Cohen, A. (1990). *A perceptual study of intonation : an experimental-phonetic approach to speech melody,* Cambridge Univ. Press.

[Ver97] Véronis *et al.* (1997). A stochastic model of intonation for French text-to-speech synthesis, *5th European Conference on Speech Communication and Technology, Eurospeech'97,* Rhodes (Greece), September 1997.

Les critères prosodiques des trois principaux types de phrases (testés sur le tchèque synthétique)

Marie Dohalská, Jana Mejvaldová

Institut de Phonétique, Faculté des Lettres, Université Charles Prague
Nám. J. Palacha 2, 116 38 Prague 1, République Tchcque
Tél.: 004202/216 19 250, fax: 004202/717 33 213
e-mail: marie.dohalska@ff.cuni.cz, jana.mejvaldova@ff.cuni.cz
internet: http://www.ff.cuni.cz

ABSTRACT

Our contribution represents another stage of the research in the area of synthesis of the Czech language. This time we concentraded on three sentence intonation types: declarative sentence, yes/no question and wh- question. The aim of the stage was to find such values of three acoustic parameters (/T/, /I/, F_0/) that they would present, if possible, an ideal compromise between the automatic prosody and the prosody which sounds natural and is adjusted by hand. In this case we disregarded possible phonostylistics variants of the model sentences. On the one hand we used the indicative values (adjusted by hand) on the yes/no question (with a change of the F_0 course, without changes of the duration and intensity values). On the other hand we used it on indicatives of other sentences with a comparable rythmical structure. Again our intention was to keep to the already discovered values. We are succeeding in finding a prevailing model of the F_0, T, I course. Generalized results of these experiments are being incorporated into the programme for the automatic prosody of the synthetic Czech. The programme is being up-dated continually.

1. INTRODUCTION

Notre communication présente une étape récente de la recherche dans le domaine de la prosodie du tchèque synthétique et décrit la démarche expérimentale conduisant à son amélioration. A partir du signal tchèque synthétique automatique (Vichová & al., 1995/96), nous poursuivons, depuis 1995, plusieurs types d'expérimentations. Au début, nous avons voulu créer au fur et à mesure une prosodie automatique "idéale", généralement valable pour le tchèque synthétique (Synthèse à Partir du Texte LPC).

Nous inspirant des recherches traitant la synthèse et la prosodie de différentes langues (surtout du français), nous nous sommes concentrés sur la méthodologie ainsi que sur les protocoles expérimentaux. Il existe bien sûr des principes valables en général pour chaque groupe de langues. Par manque des moyens techniques, les recherches équivalentes ne pouvaient pas être réalisées pour le tchèque. L'image sonore du tchèque étant très bien et minutieusement décrite d'une façon classique, plutôt descriptive qu'expérimentale [Palk92], nous reprenons l'analyse, à partir des méthodes et instruments récents. A partir de là nous élargissons la description à l'analyse de la parole spontanée afin d'établir des règles appliquables à la synthèse - ce qui est déjà fait pour de nombreuses langues [JVaiss97]. Notre idée est qu'en confrontant systématiquement l'analyse du signal de la parole naturelle avec des modèles de phrases synthétiques, nous pouvons réussir à mieux connaître le caractère sonore du tchèque standard ainsi que ses qualités typiques ou exceptionnelles. Nous espérons ainsi contribuer à l'amélioration de la qualité prosodique du tchèque synthétique et augmenter son caractère naturel. Comme matériau de base de parole naturelle, nous prenons toujours des textes continus "spontanés" enregistrés par des professionnels de locuteurs considérés comme parlant le "tchèque standard" [DohHe97]. En même temps, nous respectons une méthode déjà connue dans les années soixantes "*l'analyse par synthèse*" [BorMa67]. Ce n'est que depuis 2 ou 3 ans que nos équipements modernes nous permettent d'entreprendre des travaux beaucoup plus rapides, complexes, exacts et détaillés qu'à l'époque où les possibilités techniques ne permettaient qu'une analyse et une synthèse laborieuses, lourdes, ne permettant pas de feedback.

2. METHODE

Dans la première étape (1995/96) de nos expériences partielles, nous nous sommes attachées à créer un modèle général du début et de la fin de la phrase déclarative (de 12 syllabes au maximum) et à définir la dépendance fonctionnelle des trois paramètres acoustiques "temps"("**t**"), intensité("**i**") et fréquence fondamentale("$\mathbf{f_0}$") des phones de la syllabe initiale (régulièrement accentuée en tchèque) ainsi que dans le groupe rythmique final mono-, bi- et trisyllabique. Nous nous sommes intéressées à l'influence des changements de **t**, **i** et $\mathbf{f_0}$ sur l'acceptabilité perceptive. [Doh96].

Les résultats des expériences décrites étant connus et publiés, nous avons voulu trouver les limites perceptivement acceptables

a) des valeurs corrélées de **t** et **i**, tout en conservant les valeurs originales de $\mathbf{f_0}$;
b) ainsi que celles des 3 paramètres pris chacuns isolément.

(pour des phrases déclaratives et interrogatives de quatre à six syllabes, exprimant différentes valeurs attitudinales) [DohMej 97].

Pour atteindre le modèle de prosodie automatique le plus acceptable possible, le plus général, nous avons entammé, avec les matériaux décrits ici, le modelage des courbes de $\mathbf{f_0}$ et **i** les plus identiques que faire se peut en conservant les valeurs de durée de la phrase de départ dans toutes ses variantes (déclarative, interrogatives). Notre souci premier était de ne pas perdre le caractère naturel de la phrase de départ.

Nous avons considéré les types de phrases suivants:

- *phrase déclarative*
- *question totale* (yes-no Q)
- *question partielle (Wh-Q)*

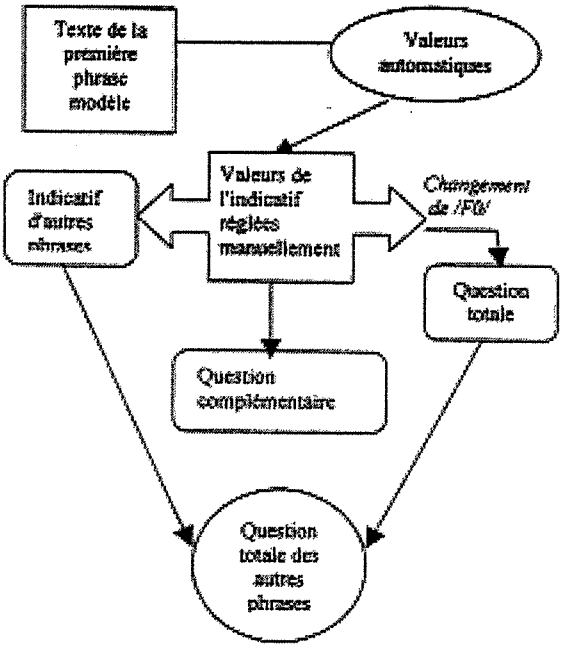


Fig. I: Relations des valeurs des phrases modèles.

Nous avons donné les valeurs de **i** et $\mathbf{f_0}$ des phrases-types aux autres phrases du corpus qui ont la même structure rythmique. Notre but est de trouver des valeurs généralement valables pour le type donné de phrase qui pourraient être utilisées dans la création du programme pour la prosodie automatique de la synthèse du tchèque - ce qui revient à dire de trouver les valeurs qui représenteraient un compromis entre l'identité absolue des valeurs (cas idéal pour l'automatisation de la prosodie) et le caractère naturel de la prosodie réglée manuellement.

En quête de valeurs optimales, nous avons considéré un corpus de 20 phrases synthétiques de 6 à 12 syllabes, déclaratives et interrogatives. Les questions totales ont été créées par changement de $\mathbf{f_0}$ uniquement, les questions partielles correspondaient sémantiquement aux phrases déclaratives. Les valeurs de /T/, /I/ et /F_0/, réglées manuellement pour la première phrase déclarative, ont été appliquées

1° aux autres phrases déclaratives;

2° aux deux autres types de phrase, avec changement de /F_0/ et maintien des valeurs de /T/ et /I/.

Pour la phrase de départ, nous avons pris un des énoncés synthétiques qui a été testé à plusieurs reprises comme le plus naturel et le plus neutre et par les sujet familiarisés avec la parole synthétique et par les sujets non familiarisés (étudiants en philologie).Nous avons essayé de conserver les proportions rythmiques des phrases de longueur différente.Le résultat recherché est d'obtenir une représentation d' un modèle général de type donné de phrase, perceptivement acceptable pour les utilisateurs pratiques de la synthèse de la parole(aide à l'accès aux textes écrits pour les non-voyants). Nous nous sommes concentrées uniquement sur les phrases "neutres", en négligeant, pour le moment, les variantes phonostylistiques possibles [DohMej 97].

3. PHRASE DECLARATIVE

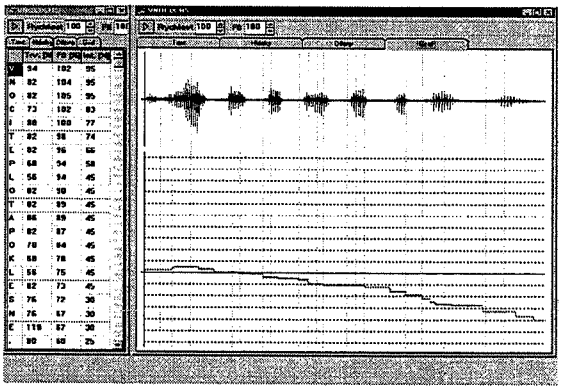


Fig. II Valeurs, signal et courbe de $\mathbf{f_0}$ (réglé manuellement) de la phrase *V noci teplota poklesne.(La température va baisser la nuit)*, selon le logiciel Víchová, Ptácek, Horák, TTS LPC 1996.

L'allure de la courbe mélodique est évidente et simple pour ce type de phrase, ainsi que le comportement des deux autres composantes **t** et **i**.

3.1 Fréquence fondamentale

F_0 s'abaisse doucement, la valeur des voyelles dans la syllabe initiale varie entre 104 et 108% des valeurs automatiques et tombe jusu'à 63 - 67% des valeurs automatiques pour la voyelle finale.
L'abaissement est régulier, un peu plus raide dès la première syllabe du dernier groupe rythmique.

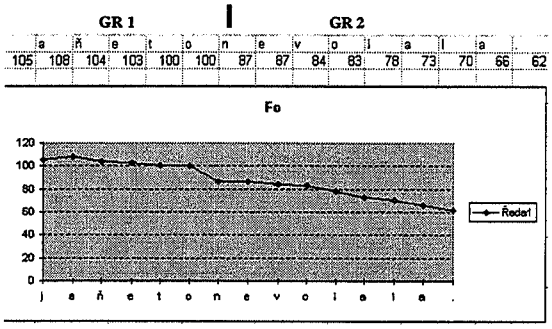


Fig.III: Courbe mélodique de la phrase à l'indicatif *Jane to nevolala. (Elle ne l'a pas téléphoné à Jeanne)*

Pour comparaison, nous montrons également ici le schéma mélodique avec le f_0 exceptionnellement augmenté au début du deuxième groupe rythmique. (conséquence de la sémantique du message - négation forte):

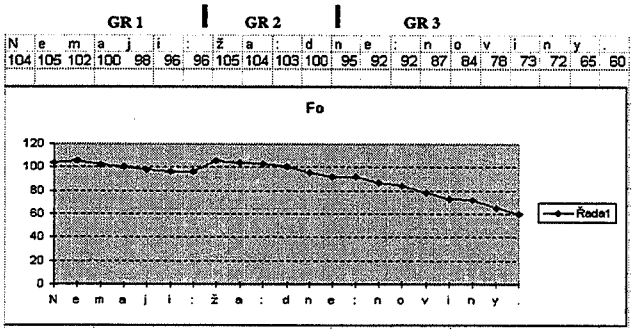


Fig.IV: Courbe mélodique de la phrase *Nemají zádné noviny. (On n'a pas de journaux).*

3.2 Intensité

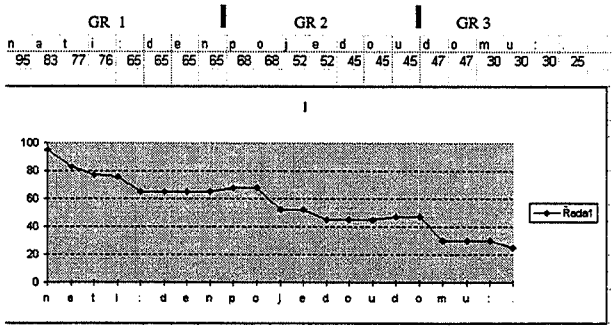


Fig. V: Intensité des phones dans la phrase *Na tyden pojedou domu (Ils vont à la maison pour une semaine.)*

En tchèque, c'est la première syllabe du groupe rythmique qui est régulièrement accentuée. Cette qualité caractéristique se reflète également dans les valeurs de l'intensité des phrases synthétiques. Les syllabes initiales des groupes rythmiques sont plus accentuées que les syllabes suivantes.

5.3 Durée

Pendant nos expérimentations, nous avons respecté la règle bien connue concernant l'allongement des voyelles à la fin absolue de la phrase. Une voyelle brève peut aller de 145% (dans une syllabe finale ouverte) à 162% (dans une syllabe finale fermée) des valeurs automatiques. Une voyelle longue (la quantité vocalique est pertinente dans le tchèque) peut atteindre 370% des valeurs automatiques Cela ne perturbe pas la tendance à l'abaissement global de l'intensité. L'intensité de la voyelle dans la syllabe initiale a été réglée manuellement à 83 ou 95% des valeurs automatiques, la voyelle dans la syllabe finale n'a que 25% de l'intensité automatique.[Doh96].

4. QUESTION TOTALE

En tchèque, c'est uniquement la mélodie qui distingue l'assertion de la question totale. Les utilisateurs de la synthèse comprennent, à partir des énoncés synthétisés, s'il s'agit d'une phrase assertive ou interrogative parce qu'un point d'interrogation à la fin d'une phrase constitue, pour le synthétiseur, une indication nette pour une montée de la mélodie (anticadence) à la fin de la phrase. Cette montée est trop rapide et la phrase est peu naturelle. Nous avons alors tenté d'atteindre une forme acceptable de la question totale juste par le changement nécessaire de f_0, **sans changer les valeurs de *t* et *I*** de la phrase déclarative.

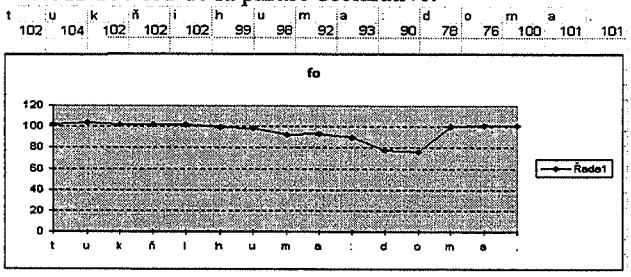


Fig. VI: Courbe mélodique de la question totale *Tu knihu má doma? (Il a ce livre à la maison?)*

On peut décrire la courbe f_0 de ce type de phrase comme un abaissement très léger, achevé par une anticadence sur la deuxième syllabe du dernier groupe rythmique (même bisyllabique). La valeur de f_0 dans la syllabe finale est pratiquement identique à celle de la syllabe initiale (102 - 108% des valeurs automatiques).

Les problèmes de modélisation de la question totale viennent de la production *naturelle* de ce type de question. La question peut être prononcée soit avec le sommet mélodique sur la dernière syllabe

(comme le montre la figure VI), soit avec le sommet sur la première syllabe du dernier groupe rythmique, les deux variantes étant orthoépiques [DohPtá94]. En comparaison des autres types de phrases modelées, le modèle synthétique de la question totale représente pour nous encore un problème à résoudre. Jusqu'à présent, nous n'avons pas réussi à modeler le deuxième "sous-type" de la question totale d'une façon perceptivement acceptable.

5 QUESTION PARTIELLE

La synthèse actuelle ne distingue pas les deux types de question.: si nous mettons un point d'interrogation à la fin de la phrase à synthétiser, la fréquence fondamentale augmente automatiquement à la fin, alors que la mélodie de la question complémentaire (comme dans de nombreuses autres langues) est plus proche de la mélodie d'une assertion (sans être vraiment identique!) que de celle d'une question totale.

5.1 Fréquence fondamentale

Nous venons de rapeler que la mélodie de ce type de phrase s'abaisse. La particule interrogative ou sa deuxième syllabe représentent son sommet. Les valeurs de $\mathbf{f_0}$ dans le dernier groupe rythmique de la question complémentaire sont identiques aux valeur de $\mathbf{f_0}$ du dernier groupe rythmique de la phrase indicative (63 - 67% des valeurs automatiques - cf. **Fig. VII**).

5.2. *Intensité*

L'intensité, quant à elle, conserve les valeurs de l'assertion dans toute la phrase et tombe de 83 - 90% de la syllabe initiale jusqu'à 25% de la syllabe finale.

6. CONCLUSION

Au cours de nos expériences, nous poursuivons un but principal: trouver des valeurs de trois paramètres acoustiques **t**, **I** et $\mathbf{f_0}$ telles qu'elles seraient perceptivement acceptables pour les trois types de phrases et pas seulement pour une phrase concrète particulière. Pour l'assertion et pour la question partielle, le modèle trouvé semble être déjà bien fidèle à la prosodie naturelle. Les valeurs ont été définies pour être introduites dans la version actualisée de la prosodie automatique du tchèque synthétique. Par le procédé mentionné au début (analyse par synthèse), nous avons rencontré un problème concernant la production du tchèque naturel: pour trouver un modèle synthétique satisfaisant de la question totale, il est nécessaire d'effectuer désormais des tests de perception de la question *naturelle* qui démontreraient lequel des deux sous-types est considéré comme plus acceptable et plus naturel, plus "neutre", et qui nous amèneraient par la suite à recommencer nos expériences de synthèse.

BIBLIOGRAPHIE

[BorMa67] Borovicková B., Malác V. (1967) *The Spectral analysis of Czech Sound Combinations*, Rozpravy CSAV, Praha

[DohPtá94] Dohalská M., Ptácek M. (1994), "Quelques remarques sur la perception du tchèque synthétique", *AUC Phonetica Pragensia VIII*, pp. 79 - 126.

[Doh96] Dohalská-Zichová M. (1996) "Testing Intelligibility at Maximum Rate", *AUC Phonetica Pragensia IX*, pp. 53 - 59

[Doh96] Dohalská-Zichová M., Dubeda T. (1996) "Rôle des changements de la durée et de l'intensité dans la synthèse du tchèque", *XXIes JEP, Avignon , Actes*, pp. 375 - 378

[DohHe97] Dohalská M., Hedánek J., (1997) "Comparing Some Parameters of Continuously Accellerated Natural Speech Signal", *Forum Phoneticum, Speech Processing*, Hans-Walter Wodarz (Ed.), Hector - Frankfurt am Main, pp. 13 - 22 .

[DohMej97] Dohalská-Zichová M., Mejvaldová J. (1997) "Où sont les limites phonostylistiques du tchèque synthétique", *16e Congrès International des Linguistes, Paris.*

[Palk92] Palkova, Z. (1992) *Fonetika a fonologie cestiny*, Karolinum, Praha

[JVaiss97] Vaissière J. (1997) "Sur la notation de la prosodie des langues", *16e Congrès International des Linguistes, Paris*

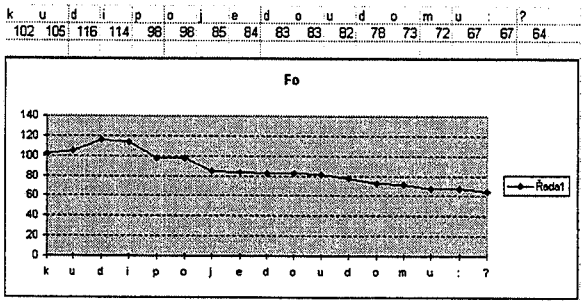


Fig. VII: Courbe mélodique de la question partielle *Kudypojedou domu? (Par où iront-ils à la maison?)*

Expériences intonatives sur l'intonation

Duvivier Delphine

Université Paris 7 Denis Diderot
Laboratoire du Coral, Université d'Orléans
Rue de Tours 45072 ORLEANS cedex 2
Tél. 02-38-41-71-06
e-mail: Institut.LETTRES@univ-orleans.fr

Abstract

Groups of listeners were asked to restitute oral texts, in which prosodic feature was manipulated. On the one hand, F0 neutralization affects syntactic structures. On the other hand, anticipated stress affects grammatical units, contrastive stress affects lexical units.
Considering the results of this experiment, it appears that, when F0 is no more effective, listeners realize restitutions which are quantitatively less performant. Dealing with the two types of stress, it doesn't make the access to information in oral texts easier. Though acoustic parameters variate, the restitution of stressed unit is executed thanks to its semantic value.
Actually, every unit doesn't have the equal compulsion of restitution.

1.Introduction

Notre propos est d'examiner si F0 et l'accent jouent un rôle dans le balisage et la restitution des unités du lexique, de la structure et de l'information par un auditeur. Cette recherche se nourrit de plusieurs présupposés théoriques quant au rôle de l'intonation dans la structuration de la langue. Les travaux de Philippe Martin [Mar75] ont démontré la fonction de corrélation de l'intonation avec les caractéristiques syntaxiques de l'énoncé: cette corrélation se servant notamment d'indices contenus dans la mélodie de la phrase; *les éléments suprasegmentaux sont donc interprétés dans un rôle d'indicateur de la structure syntaxique de la proposition* (la phrase et la proposition étant respectivement l'expression et le contenu de l'énoncé). Mary-Annick Morel [Mor92] travaille sur les faits de thématisation corrélés à l'intonation: elle montre que les hauteurs mélodiques -les points cruciaux de hauteur sont mesurés à la syllabe finale et au point d'attaque d'un groupe rythmique- apportent une information sur la nature de la relation qu'entretiennent deux groupes se suivant: autonomie, emboîtement ou rupture. Daniel Hirst [Hi97] propose une interprétation de l'intonation dans le cadre de la Théorie de Pertinence. En effet, cette théorie fournit un cadre naturel, selon Hirst, capable d'expliquer comment l'intonation d'un énoncé va contribuer à son interprétation syntaxique, sémantique et pragmatique. F0 peut également jouer un rôle dans l'expression de l'émotion et de l'attitude en les spécifiant : on se référera ici aux travaux de P. LEON [Leo93], dégageant notamment des contours mélodiques, des valeurs d'intensité et de durée idiomatiques de certaines émotions, telles la tristesse ou la colère. De même, F0 peut être le critère désambiguïsateur lorsque un énoncé du type *un joueur de football américain* offre une double interprétation; ainsi, une identité de contour descendant sur *joueur* et *américain* marquera l'identité de la relation syntaxique: l'adjectif qualifie *joueur*.
Quant aux études sur l'accentuation en français, elles mettent en évidence le rôle des indices acoustiques dans la perception des syllabes accentuées (Rigault, 1962 [Ri62]; Rossi et alii, 1981 [Ro81]). En outre, on a pu distinguer des spécificités d'utilisation des indices acoustiques pour chaque langue; à l'instar des individus qui mettent en place des stratégies diverses pour la perception, notamment, de l'accent.

2.Problématique et hypothèses

Si les indices acoustiques permettent la perception de l'accent en français, les travaux de Fonagy [Fo79] émettent l'hypothèse que d'autres facteurs, sémantique par exemple, interviennent dans les stratégies de perception. Quant aux variations de mélodie, elles indiquent la structure syntaxique de l'énoncé: qu'en est-il lorsqu'elle n'en est plus un indice?
les hypothèses qui sous-tendent ce travail sont les suivantes: (a) lorsque F0 est neutralisé, obtenant ainsi du recto-tonal, les marques de la structure syntaxique sont appauvries: la structure informationnelle l'étant de ce même fait, l'auditeur devrait alors avoir des difficultés à repérer l'information; (b) l'accent porté sur des unités non accentogènes (catégories fonctionnelles: clitiques, qui ne sont pas susceptibles en français de recevoir un accent) impliquera une perturbation à l'oral pour la perception de l'information par un auditeur; (c) l'accent porté sur des unités lexicales permettra une meilleure perception de l'information, puisque l'on fera coïncider accent et unité accentogène (unité lexicale).

3.Méthodologie

3.1.Tests de perception

Test 1: il portera sur F0. Trois administrations sont réalisées: un document est lu; un document est donné à l'oral avec une structure intonative manipulée par neutralisation de F0 (neutralisation par voix naturelle et non par synthèse de la parole; c'est pourquoi une augmentation du débit de parole a été observé, entraînant ainsi une diminution des pauses): c'est une structure recto-tonale; un document est donné à l'oral avec une structure intonative non manipulée. On aura donc des sujets auditeurs et des sujets lecteurs.
Test 2: un accent est porté à l'oral sur des unités à propriété sémantique référentielle: il s'agit donc de rajouter au schéma accentuel préalablement existant des accents de type contrastif; cet accent souligne le choix particulier qui est fait dans une unité d'un élément d'un paradigme donné. Une autre version orale ne rajoutant aucun accent au schéma accentuel sera proposé. Cette seconde version permettra le contrôle de la première.
Test 3: un accent d'insistance, qui assure la mise en relief d'une unité, va être porté sur des unités clitiques suivies immédiatement à droite par une unité pleine. La même administration que le test 2 sera réalisée ici.

3.2. Sujets

Les 48 sujets passant les trois tests sont étudiants à l'Université d'Orléans en droits, lettres, langues, sciences humaines. Agés de 18 à 27 ans, ils sont francophones natifs et d'origines socio-culturelle similaires.

3.3.Matériel et procédure

Les sujets lisent (uniquement pour T1) ou écoutent des documents relatant des expériences scientifiques: la consigne étant, à la suite de cette écoute ou lecture, d'effectuer un rapport complet sous forme de rappel libre, c'est-à-dire redonner le plus grand nombre d'éléments possibles sans tenir compte de l'ordre de présentation de l'information dans le texte. Le caractère scientifique des documents a été choisi de façon à ce que la restitution ne soit réalisée que relativement à des connaissances internes à l'expérience et non externes, par l'intermédiaire des connaissances des sujets: on rappellera qu'il n'y a aucun étudiant en sciences. Quant aux temps des documents de base à partir desquels sont réalisés les rapports (ou restitutions), il a été décidé qu'ils ne devaient pas excéder 1mn, 1mn 30. En effet, il nous semble que cette durée permettra de lire ou d'écouter des textes dont la structure argumentative et informationnelle est suffisamment riche.

4. Résultats

On procédera par relevés lexicaux -substantif, adjectif, verbe et adverbe- intégrant la composante référentielle, *le quelque chose dont on parle*, définie par R. Jakobson [Jak63] dans son schéma de la communication.

4.1.T1

Par l'établissement des moyennes établies pour chaque catégorie syntaxique -substantif, adjectif, verbe et adverbe (et nom propre uniquement pour T1)-, on obtient le résultat suivant (figure 1). On fera également apparaître le taux de restitution pour la totalité des catégories syntaxiques, c'est-à-dire le taux de restitution donné pour une version. Il en sera de même pour T2 et T3. On précisera que le taux de restitution est calculé par la mise en rapport, pour une catégorie syntaxique, du nombre d'unités restituées -reprises du document original- et le nombre total d'occurrences produits: unités restituées pour une catégorie/total d'occurrences produites pour cette même catégorie.

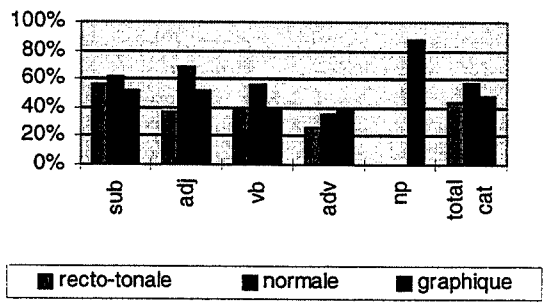


Figure 1

Interprétation: T1

La figure 1 donne les taux de restitution de 5 catégories syntaxiques pour les 3 versions analysées: recto, normale et graphique. On pourrait ainsi supposer que l'hypothèse de départ se confirme: lorsque les variations mélodiques se réalisent, le taux de restitution des auditeurs est plus important avec 56.9% et une moyenne de 19: les substantifs, adjectifs et verbes sont restituées substantiellement. Ce qui n'exclut pas les variations internes, puisque l'on observe, notamment pour les noms propres, qu'ils ne sont restitués qu'à la suite de la lecture: aucune unité de ce type n'est redonnée après une écoute normale ou recto-tonale.
On pourrait donc penser que si la version normale offre une restitution importante, c'est parce que F0 joue effectivement un rôle de marqueur syntaxique. La structure de l'énoncé étant indiquée, la structure informationnelle, établie, entre autre, par l'intermédiaire des connecteurs, est marquée prosodiquement et perçue par l'auditeur. En conséquence, à la consigne *redonner tout ce dont vous vous souvenez à propos de ce document écouté*, les auditeurs réalisent des restitutions

performantes. Lorsque ces variations de F0 sont neutralisées (la version recto donne une moyenne de 12 et un taux de 43.6%), le taux de restitution est quantitativement plus faible; on peut dire que les variations de hauteur n'étant plus effectives et la durée étant elle-même modifiée -ne marquant plus que faiblement la structure syntaxique de l'énoncé-, il manque des indices de mélodie et de durée à l'auditeur pour repérer la structure informationnelle de l'énoncé, qui est alors appauvrie. La version graphique avec un taux de restitution de 48.5% et une moyenne de 20 corrobore les interprétations données pour la version recto. Ceci confirme ainsi la fonction de corrélation entre la mélodie de la phrase et son organisation syntaxique, qui permet en conséquence à l'auditeur de repérer l'information, l'organisation informationnelle du discours à l'oral. F0 joue, à la lumière de cette expérience, un rôle dans l'accès à l'information à l'oral; et ce, par l'intermédiaire de la composante syntaxique.

4.2.T2

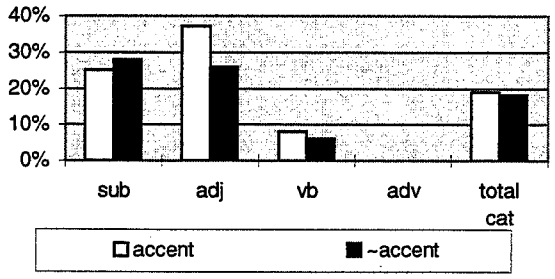


Figure 2

Les chiffres avancés ici concernent la restitution des unités accentuées pour chaque catégorie syntaxique et version. On observe que lorsque l'accent contrastif est rajouté au schéma accentuel, substantifs, adjectifs et verbes n'ont pas une restitution plus importante: les taux sont identiques à la version *~accent*; de même, les adverbes sont donnés identiquement, c'est-à-dire aucun (figure 2). On peut donc voir que malgré la fonction de mise en relief de cet accent contrastif, il n'a pas d'incidence sur la restitution des unités qui vont le porter: les auditeurs ne semblent pas le percevoir comme un indice de récupération; les taux de restitution des versions *accent* et *~accent* sont respectivement de 18.5% et 17.5% et leur moyenne de 16.5 à toutes deux. Toutefois, si l'on observe les unités accentuées, on remarque ce qui suit (on précisera que le document de base relatait l'intérêt que les psychologues ont à replonger les bébés dans l'eau dès leur plus jeune âge): quelle que soit la version mise en place, les mêmes unités sont toujours restituées plus fortement. Ainsi, on trouvera toujours *liquide*, *psychologues*, *aisance*, *élément* et *nager*. De même, les unités peu ou pas restituées sont identiques: *jamais*, *aquatique*, *tendre*, *primordial*. Si l'on retrouve systématiquement les mêmes unités plus ou peu ou pas restituées, on pourrait confirmer les premiers résultats énoncés précédemment en terme de restitution: l'accent contrastif ne joue pas de rôle, à l'oral, dans l'accès à l'information, malgré la perception des variables acoustiques par les auditeurs.

4.3.T3

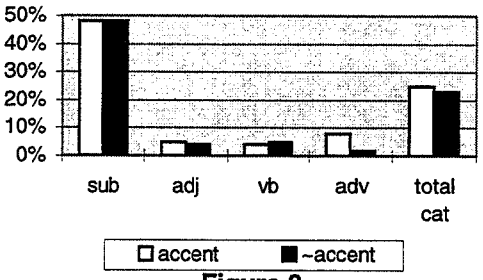


Figure 3

Les données chiffrées sont établies identiquement que dans T2. Cet accent anticipé portant sur des items non accentogènes est susceptible, de ce fait, d'introduire une perturbation dans la perception de la chaîne sonore. Or, on observe en figure 3 que les taux de restitution pour chaque catégorie syntaxique sont identiques: pour les substantifs, adjectifs, verbes et adverbes, aucune différence significative n'est observable. Quant aux taux établis pour chaque version, on note qu'ils sont également similaires: 25.3% pour *accent* (moyenne: 20.5) et 23.3% pour *~accent* (moy.: 22.2). Comme précédemment pour T2, nous nous intéressons aux unités accentuées dans le document de base (on précisera que ce texte concernait la respiration du foetus en situation intra-utérine): il semblerait que des commentaires identiques à l'accent contrastif puissent être réalisés. En effet, les mêmes constantes se dégagent: quelle que soit la version -avec ou sans accent-, on retrouve systématiquement les mêmes unités plus fortement redonnées et les mêmes unités en quantité moindre ou nulle. *Foetus*, *mère*, *placenta*, *poumons*, *cordon* et *liquide* ont des restitutions comparativement plus importantes alors que *pompe*, *placentaires*, *nouveau-né*, *intermédiaire*, par exemple, ont, elles aussi, des restitutions plus faibles voire nulles et ce, encore une fois, quelle que soit la version.

Interprétation: T2 et T3

On a donc deux types distincts d'accent. L'un, contrastif, qui souligne le choix particulier d'une unité. Or, les auditeurs ne s'en servent pas comme un indice, malgré la consigne préalable à l'écoute du document. Qu'il soit rajouté ou non au schéma accentuel, les auditeurs ne restituent les unités qu'en fonction de leur valeur sémantique. L'autre, anticipé, qui pouvait introduire une perturbation dans le déroulement de la chaîne sonore. A l'instar de l'accent contrastif, le traitement perceptif de cet accent n'implique pas une restitution moins importante.

Elle sera identique à la restitution réalisée sans l'ajout de cette marque. La valeur sémantique d'une unité sera le même critère pertinent et déterminant pour qu'un item du document original soit redonné par les auditeurs. En conséquence, lors de ces tests de perception, l'accent, contrastif ou anticipé, ne sert pas le traitement perceptif de l'information: il ne permettra pas à une unité d'être plus restituée, comparativement à une version où un ajout de ce type au schéma accentuel existant est réalisé. Seule la valeur sémantique des unités a joué un rôle ici dans l'accès à l'information. Toutefois, dans ces tests, le poids accentuel des unités n'est jamais modifié ni dans T2, ni dans T3: il reste identique. Or, les travaux de I. Fonagy [Fo80] démontrent la corrélation entre poids accentuel et poids sémantique: dans ces expériences, il apparaît que *l'accent réagit à des divergences d'ordre sémantique; on pourrait dire que le facteur qui a le plus d'incidence sur la place de l'accent se situe au niveau du message en se fixant sur l'élément qui possède un rapport d'information plus important.*[...] *L'adjectif exerce une plus forte attraction sur l'accent que le substantif. Dans un syntagme attributif dissyllabique c'est généralement l'adjectif monosyllabique qui porte l'accent, même si le substantif précède l'adjectif. le poids accentuel semble correspondre à un degré d'intérêt plus élevé, qu'aura, en principe, le qualificateur en face du qualifié.* [...] *On pourrait donc supposer que le poids accentuel reflète, toutes conditions égales par ailleurs, le poids sémantique d'un mot ou d'une classe de mots.*

5.Conclusion

Ces expériences perceptives sur l'intonation montrent que F0 et l'accent, un paramètre et un phénomène suprasegmentaux, ne jouent pas des rôles identiques dans l'accès à l'information dans le discours oral. Le critère de la variation sera ici déterminant. Variant et étant neutralisé, F0 appauvrit les marques de la structure syntaxique de l'énoncé: de ce fait, il appauvrit la structure informationnelle de ce même énoncé; les variations de hauteur mélodique ne se modifiant pas, l'auditeur n'accède plus à la composante informationnelle parce qu'il n'accède plus à la composante syntaxique. Quant à l'accent, son "poids" n'est pas modifié non plus: il sera identique pour une unité qui a ou non un rapport d'information important dans l'énoncé qui l'intègre; ainsi, il n'aura pas d'incidence sur la restitution d'une unité. Et ce, à l'inverse des travaux de Fonagy qui montrent la corrélation entre poids accentuel et poids sémantique: le second affectant l'importance des variations prosodiques du premier.

Bibliographie

[Fo80] FONAGY I. (1980), "L'accent en français contemporain: accent probabilitaire", *Studia Phonetica*, L'Accent en français contemporain, n°15, Didier, Paris.

[Hi97] HIRST D. (1997), "Interprétation de l'intonation dans le cadre de la Théorie de Pertinence", XVIè Congrès International des Linguistes, 20-25 juillet 1997, CNRS LLACAN, Paris.

[Jak63] JAKOBSON R. (1963), *Essais de linguistique générale*, Ed: Minuit, Paris.

[Leo93] LEON P. (1993), *Précis de phonostylistique. Parole et expressivité*, Ed. Nathan U.

[Mar75] MARTIN Ph. (1975), "Intonation et reconnaissance automatique de la structure syntaxique", *8èmes JEP*, Toulouse, 28-30 mai 1975.

[Mar80] MARTIN Ph. (1980), "Une théorie syntaxique de l'accentuation en français", *Studia phonetica*, L'Accent en français contemporain, n°15, Didier, Paris.

[Mor92] MOREL M.-A. et RIALLAND A. (1992), "Autonomies, emboîtements et ruptures dans l'intonation française", in Actes du Colloque CERLICO, *La subordination*, Presses Universitaires de Rennes II.

[Ri62] RIGAULT A. (1962), "Rôle de la fréquence, de l'intensité et de la durée vocalique dans la perception de l'accent en français", in *Proceedings of the 4th International Congress of Phonetics Sciences*, La Haye, Mouton, 1961, pp. 735-748.

[Ro81] ROSSI M., HIRST D., MARTIN Ph. et NISHINUMA Y. (1981), *L'intonation: de l'acoustique à la sémantique*, Paris, Klincksieck.

Pondération d'indices prosodiques dans un système automatique de prédiction syntaxico-rythmique d'un énoncé

†*Philippe Langlais and* ‡*Jean-Luc Cochard*

†Centrum för talteknologi (CTT) – KTH
Box 70014 — S-10044 Stockholm
Tél.: ++46 8 790 75 65
Fax: ++46 8 790 78 64
http://www.speech.kth.se/

‡CT-SPI, Swisscom
Ostermundingenstr. 93 — CH-3000 Berne
Tél.:++41 31 342 87 34
Fax: ++41 31 342 29 53
Jean-Luc.Cochard@Swisscom.com

ABSTRACT

This paper concerns the study of information derived from the melodic, temporal and intensity characteristics of the material to be recognized in a speech recognition system, in French.

More precisely, it describes experiments we achieved at the suprasegmental levels with a system that outperform automatic correlation between prosodic labels and linguistic organization of a message to decode. An overview of the system is described along with the results of experiments carried out to determine which prosodic indexes are best-suited for syntactic and rhythmycal prediction.

1. INTRODUCTION

Il est bien connu que les structures prosodiques et syntaxiques mises en œuvre dans le processus de communication ne sont ni identiques ni non corrélées. Savoir quand et comment ces deux structures sont mises en relation est d'un intérêt évident dans des tâches telles que la désambiguïsation d'hypothèses concurrentes dans un système de dialogue. Ce type d'application explique en partie le regain d'intérêt récent de l'intégration nécessaire de la prosodie en reconnaissance automatique de la parole [2, 6, 7, 4].

Cet article présente des expériences visant à répondre à ces deux questions :

- Est-ce que l'information prosodique est, lorsque qu'extraite automatiquement, suffisante à la prédiction de la structure syntaxique et/ou rythmique d'un énoncé inconnu ?
- Dans l'affirmative, existe t-il des indices prosodiques, parmi l'ensemble de ceux utilisés, qui jouent un rôle prépondérant dans le processus de prédiction ?

2. BASE DE DONNÉES

Les deux expériences décrites ici utilisent un sous-ensemble d'une base de parole téléphonique réalisée à l'IDIAP en collaboration avec Swisscom [1].

2.1. *Le corpus de phrases*

Un corpus d'apprentissage de 500 phrases a été sélectionné. Prononcées par 50 appelants (via différentes lignes téléphoniques), ces phrases sont des répétitions en nombre inégal de 80 phrases différentes, de structure syntaxique simple contenant de 4 à 17 voyelles (voir figure 1).

⋆ Le tout apparaît dans un bilan clair.
⋆ Cette révélation tardive souligne la gravité du malaise.

Figure 1: Exemples de phrases utilisées.

À des fins de tests, un corpus de 300 phrases a aussi été sélectionné. La plupart de ces phrases étant des répétitions (généralement par d'autres locuteurs), des phrases du corpus d'apprentissage.

2.2. *Le corpus de nombres décimaux*

Un corpus d'apprentissage de 500 nombres décimaux (voir figure 2) enregistré dans les même conditions (lignes téléphoniques, 50 locuteurs, etc.) a été reuni.

⋆ Treize mille deux cent quarante virgule dix.
⋆ Quatre mille virgule huit.

Figure 2: Exemples de nombres décimaux utilisés.

Enfin, un corpus de test de 280 nombres décimaux a été collecté. 148 de ces nombres ont une structure "syntaxico-rhytmique" apparaissant au moins une fois dans le corpus d'apprentissage.

3. ÉTIQUETAGE PROSODIQUE

En dépit de systèmes d'étiquetage émergant comme ToBI [8], l'étiquetage prosodique n'a toujours pas trouvé de solution satisfaisante. Un ensemble d'étiquettes a été retenu dans cette étude (9 indices de durées, 9 d'intensité et 22 indices décrivant la fréquence fondamentale). Chaque indice, de nature binaire (présent/absent), décrit chaque noyau vocalique (émergence locale d'un paramètre au regard des voyelles adjacentes, émergence globales, extrema, etc.). Ce processus simple et complètement au-

tomatique [5] calcule pour un signal de parole donné une matrice prosodique qui constitue la seule entrée du processus de reconnaissance que nous décrivons maintenant.

4. LE SYSTÈME DE PRÉDICTION

4.1. *Hypothèse de travail*

Le but des expériences présentées est de valider ou d'infirmer l'hypothèse de travail suivante : la distribution des configurations des paramètres prosodiques sur l'ensemble d'une phrase n'est pas aléatoire, mais répond bien au contraire à des contraintes structurelles (syntaxiques, rythmiques, sémantiques) suffisamment régulières pour un traitement automatique.

4.2. *Description générale*

Le système décrit ne fait aucune hypothèse quant à l'organisation des différentes contraintes régissant les paramètres prosodiques pas plus qu'il ne présuppose l'existence d'entités prosodiques particulières comme par exemple *l'accent*, notion pour laquelle nous n'avons pas trouvé de caractérisation acoustique "robuste". L'approche fondatrice du système est une étude menée par les données des corrélations entre différents niveaux linguistiques et les indices prosodiques mesurés automatiquement. Le système est composé de deux étapes décrites ci-après.

4.3. *Phase d'apprentissage*

Les étiquettes prosodiques calculées depuis un signal de parole, ainsi qu'un arbre syntaxique associé (dont les feuilles sont alignées au signal) constituent les entrées du processus d'apprentissage. Le formalisme descriptif des arbres syntaxiques respecte les règles suivantes où α et β représentent des symboles non terminaux (définis par l'utilisateur) :

$\alpha(\beta, n)$	i.d. α est un groupe de n voyelles décrit par β,
$\alpha(\epsilon, n)$	i.d. α est une feuille de n voyelles,
$\alpha(\beta, \varepsilon)$	i.d. α est un groupe décrit par β sans spécification du nombre de voyelles,
$\alpha.\beta$	i.d. α et β,

Un exemple d'arbre est reporté en figure 3a. La phase d'apprentissage consiste à créer un graphe décrit de la manière suivante :

- chaque nœud (p) contient la caractérisation prosodique d'une structure syntaxico-rythmique ($\mathcal{S}_p$),

- chaque arc (a_{ij}) représente une contrainte (syntaxique ou rythmique) affinant la description de la structure décrite par le nœud i.

Les arcs sont déterminés à partir des arbres syntaxiques disponibles dans le corpus d'apprentissage en appliquant, niveau par niveau, les règles de généralisation suivantes (voir figure 3) :

a- un groupe α de n voyelles décrit par β est un groupe α de n voyelles ;

b- un groupe α de n voyelles décrit par β est un groupe de n voyelles ;

c- un groupe α de n voyelles est un groupe α ;

d- un groupe α de n voyelles est un groupe.

$$P = \begin{bmatrix} a- & \alpha(\beta,n) & \rightarrow & \alpha(\epsilon,n) \\ b- & \alpha(\beta,n) & \rightarrow & \mathcal{G}(\epsilon,n) \\ c- & \alpha(\beta,n) & \rightarrow & \alpha(\epsilon,\varepsilon) \\ d- & \alpha(\beta,n) & \rightarrow & \mathcal{G}(\epsilon,\varepsilon) \end{bmatrix}$$

où $\mathcal{G}$ est un symbole désignant un groupe quelconque. Une observation (i.e. un arbre syntaxique) de profondeur p peut générer jusqu'à $\frac{p\times(p+3)}{2}$ nœuds dans le graphe d'apprentissage. En pratique, le nombre de recouvrements (dépendant de l' application visée) limite la croissance du graphe.

Chaque nœud de ce graphe tient le compte des indices prosodiques localisés au début ou à la fin de chaque feuille de la structure décrite par le nœud. Chaque structure syntaxico-rythmique est ainsi décrite par une matrice prosodique de n lignes (n étant le nombre d'indices considérés : 40 ici) et de k colonnes ; k étant borné par le double du nombre de feuilles de la structure décrite.

À la fin du processus d'apprentissage (complètement décrit dans [5, pp. 134–138]) un utilisateur peut interroger le système afin d'obtenir des informations comme des contours paramétriques moyens associés à une structure donnée de phrase, ou encore des statistiques sur les points de rendez-vous entre la syntaxe, le rythme et la prosodie.

a) $PH($ $SS(GN(Art(\epsilon,1).NC(\epsilon,3),4),4).$
$SV(VB(\epsilon,1).GN(ART(\epsilon,1).$
$NC(\epsilon,1).ADJ(\epsilon,1),3),5),9)$

b) 1- $PH(SS(\epsilon,4).SV(\epsilon,5),9)$,
2- $PH(\mathcal{G}(\epsilon,4).\mathcal{G}(\epsilon,5),9)$,
3- $PH(SS(\epsilon,\varepsilon).SV(\epsilon,\varepsilon),9)$,
4- $PH(\mathcal{G}(\epsilon,\varepsilon).\mathcal{G}(\epsilon,\varepsilon),9)$

Figure 3: a) Structure syntaxico-rythmique de la phrase : "Une pique-niqueuse mange une pomme verte" avec les symboles utilisateur suivants : PH phrase, SS groupe sujet, GN groupe nominal, ART article, NC nom commun, SV groupe verbal, VB verbe, ADJ adjectif. **b)** nœuds crées par l'application des règles de production au niveau 2 de la structure décrite en a).

4.4. *Prédiction structurelle*

Ce stage consiste à émettre des hypothèses sur la structure syntaxico-rythmique d'un énoncé, étant donné la matrice prosodique le caractérisant (M_o qui est une matrice $n \times l_o$, l_o étant le nombre présumé de voyelles dans le signal) et un graphe d'apprentissage (G). Ce processus met en jeu un parcours de graphe dont les points importants sont rappelés ici (voir [5, pp. 141–144] pour plus de détails).

Soit p un nœud appartenant à G, décrivant une structure ($\mathcal{S}_p$) de f feuilles ; une matrice prosodique $n \times l_p$ (avec $f < l_p < 2 \times f$) lui est associée. Les deux étapes du parcours de graphe sont les suivantes :

réduction de M_o par $\mathcal{S}_p$: étape qui consiste à trouver un ensemble de f couples (d_i, f_i) de colonnes de M_o, correspondant à un découpage possible du signal de parole en f constituants (ceux décrits par $\mathcal{S}_p$). Les conditions suivantes sont également vérifiées : pour chaque $i \in [1, f]$, $d_i = f_i + 1$, $f_i \geq d_i$, $d_1 = 1$, $f_{l_p} = l_o$. Les matrices résultantes sont notées M_{o_p}.

mesure de similarité : entre deux matrices M_{o_p} et M_p (de même dimension) par le calcul de $d(M_{o_p}, M_p)$:

$$\begin{cases} d(M_{o_p}, M_p) & = \dfrac{\sum_{i=1}^{n} \left(\alpha_i \sum_{j=1}^{l_p} \delta_{ij} \right)}{\sum_{i=1}^{n} \alpha_i} \\ \text{et } \delta_{ij} & = \begin{cases} M_p(i,j) \ si \ M_{o_p}(i,j) = 1 \\ 0 \ sinon \end{cases} \\ \text{et } \alpha_i & \text{poids du } i^{eme} \text{ indice} \end{cases}$$

La valuation d'une hypothèse (potentiellement tout nœud de G) est finalisée en retenant le meilleur chemin de la racine de G au nœud considéré. Le score associé à un chemin particulier est la moyenne des scores de similarité entre les matrices M_{o_q} et M_q, pour chaque nœud q du chemin.

5. Expériences

5.1. *Protocole de test*

Dans ces deux expériences, nous avons imposé à notre système de réaliser une prédiction complète de la structure syntaxico-rythmique d'un énoncé inconnu ; c'est à dire de déterminer un arbre syntaxique complet ainsi que la spécification du nombre de voyelles de chacune de ses feuilles. Seule la matrice prosodique calculée à partir du signal de parole est utilisée.

Le nombre moyen de réponses possibles était de 15 aussi bien pour les phrases que pour les nombres décimaux.

5.2. *Expérience A*

Dans cette expérience, nous avons fait l'hypothèse que chaque indice prosodique jouait un rôle identique dans le processus de reconnaissance (i.e. $\alpha_i = 1$ pour tout indice i).

Phrases Le système montre des aptitudes intéressantes : sur le corpus d'apprentissage plus de 90% des hypothèses formulées en première position sont les bonnes (figure 4a). Sur le corpus de test, ce taux avoisine 60% (4b). Ces résultats peuvent être positivement comparés à ceux obtenus par un score aléatoire (figure 4c).

Nombres décimaux Là encore, le système présente un bon taux de reconnaissance (plus de 80%, voir la figure 4d). Les performances sur le corpus de test réduit aux nombres dont la structure a déjà été vue lors de l'apprentissage sont moins bonnes (proches de 50%, voir la figure 4e). L'analyse précise des hypothèses a cependant montré que même en cas d'erreur, le système était fiable quand à la prédiction de l'emplacement du mot "virgule" (figure 4f).

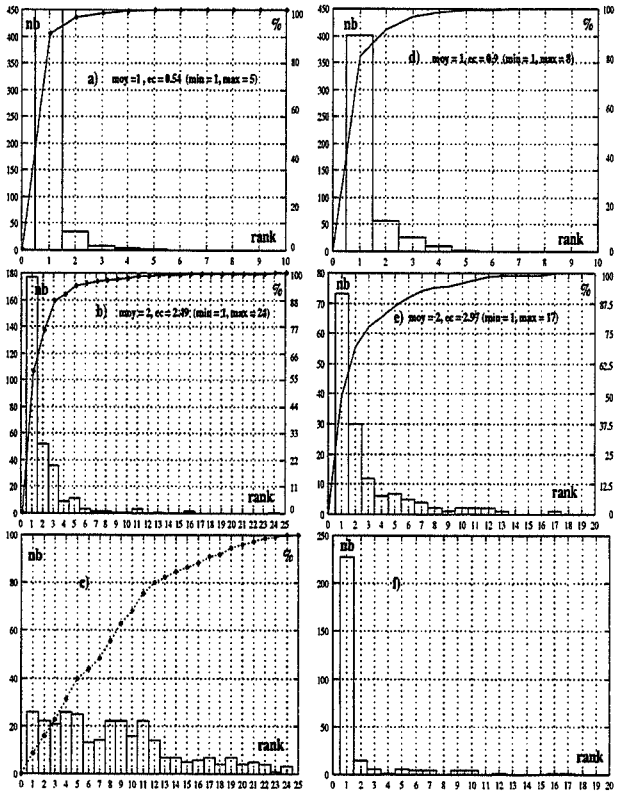


Figure 4: L'axe des abscisses indique le rang dans la cohorte d'hypothèse de la bonne solution. L'axe des ordonnées à gauche indique le nombre d'hypothèses réalisées et à droite, le pourcentage cumulé de succès à un rang donné. **a)** prédictions sur le corpus d'apprentissage. **b)** prédictions sur le corpus de test (300 sentences). **c)** prédictions sur le corpus de test à l'aide d'une mesure de similarité aléatoire. **d)** prédictions sur le corpus d'apprentissage des nombres. **e)** prédictions sur le corpus de test des nombres. **f)** prédiction de l'emplacement du mot "virgule" dans les nombres du corpus de test.

Nous avons essayé par un ensemble de règles mises au point sur le corpus d'apprentissage de décrire prosodiquement le mot "virgule" et son contexte droite et gauche immédiat. Cet ensemble de règle appliqué au corpus de test n'a jamais permis d'approcher les performances de notre système. Ceci tend à confirmer le fait que l'étude de l'information prosodique sur l'ensemble d'un énoncé (phrase, nombre, mot, etc.) semble plus pertinente que l'analyse locale des phénomènes prosodiques [9]. Ceci illustre une des limitations importantes de l'intégration efficace de la prosodie dans les systèmes de reconnaissance actuels.

5.3. *Expérience B*

Deux expériences supplémentaires ont été réalisées afin de déterminer les indices prosodiques les plus pertinents dans notre système. Dans la première (B1), nous avons évalué séparément la performance de chaque indice pour la tâche de prédiction. Dans la seconde expérience (B2), nous avons tenté de définir la combinaison optimale (au regard des performances sur nos corpus d'apprentissage) des poids (α_i) associé à chaque indice prosodique i.

Dans les deux expériences nous avons utilisé une fonction d'erreur E où r_i est le rang de la solution dans la cohorte d'hypothèses formulées et N le nombre de tests.

$$E = \frac{\sum_{i=1}^{i=N}(r_i - 1)}{N}$$

Expérience B1 Nous reportons quelques observations faites à partir des résultats obtenus sur la tâche de prédiction des structures complètes des phrases du corpus de test.

a) Prendre en considération tous les indices avec un poids égal à l'unité donne de meilleurs résultats qu'en ne considérant qu'un sous-ensemble des indices.

b) En regardant les taux d'erreur mesurés lorsque les prédictions sont réalisées en ne considérant les indices que d'un seul paramètre, on constate que les indices de durée sont les plus pertinents, suivis par les indices d'intensité. Les étiquettes prosodiques caractérisant le minimum et le maximum de chaque paramètre sont les indices prosodiques les plus bruités.

c) Plus précisemment, l'allongement de la durée de la voyelle (plus de 20ms au dessus de la durée moyenne des voyelles sur la phrase considérée) est l'indice le plus pertinent. Les indices de pente caractérisant le paramètre de fréquence fondamentale (montant, descendant, plat) sont ensuite les plus utiles.

Expérience B2 Dans cette expérience, nous avons supposé que E était une fonction dépendant seulement des n variables $\alpha_1, \alpha_2, ..., \alpha_n$ (i.e. les poids prosodiques). Nous avons alors utilisé la méthode d'optimisation non dérivative du *Simplex* [3] afin de trouver automatiquement l'ensemble de coefficients minimisant la fonction d'erreur E.

Les poids ainsi obtenus confirment les observations faites dans l'expérience B1. Le taux de performance mesuré sur la tâche de prédiction des structures des phrases du corpus de test a ainsi été augmenté de 0.8% (gain proche d'un rang en moyenne).

6. CONCLUSION

Le système présenté ici permet de prédire la structure syntaxico-rythmique d'un signal de parole à partir des seules informations prosodiques calculées automatiquement. Les premières expériences (tous les indices prosodiques jouant un rôle identique dans la fonction de notation) indiquent pour les phrases des performances de 90% sur le corpus d'apprentissage, de 60% sur le corpus de test et de moins de 10% avec un score aléatoire. Ceci permet donc de valider notre hypothèse de travail et de démontrer que notre système d'étiquetage prosodique est pertinent.

Dans les expériences visant à déterminer l'information la plus utile, nous notons que l'indice d'allongement de la durée des voyelles est l'indice le plus pertinent. Nous rejoignons en cela les observations faites dans des études antérieures [10]. D'autres expériences sur des corpus de taille plus importante sont cependant nécessaire.

REMERCIEMENTS

Nous tenons à remercier l'institut IDIAP pour la base de données utilisée dans cette étude ainsi que Pierre Jourlin pour ses précieux conseils.

BIBLIOGRAPHIE

[1] G. Chollet, J-L. Cochard, A. Constentinescu, C. Jaboulet, and P. Langlais. *Swiss-French Polyphone and PolyVar : Telephone Speech Databases to develop Model Inter- and Intra-speaker Variability*, chapter 8, pages 117–135. CSLI, 1997.

[2] A. Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *ICASSP*, volume II, pages 169–172, Adelaïde (Australia), 1994.

[3] Nelder J.A. and Mead R. A simplex method for function minimization. *Comput. J.*, pages 308–313, Janv. 1965.

[4] Hirose K. *Disambiguating Recognition Results by Prosodic Features*, volume 19, chapter 21, pages 327–342. Springer-Verlag, 1997.

[5] P. Langlais. *Traitement de la prosodie en reconnaissance automatique de la parole*. PhD thesis, Université d'Avignon, 1995.

[6] P. Langlais and J-L. Cochard. The use of prosodic agents in a cooperative automatic speech recognition system. In *International Congress of Phonetic Sciences*, volume 4, pages 292–295, Stockholm, Sweden, August 13–19 1995.

[7] Ostendorf M. and Ross k. *A Multi-level Model for Recognition of Intonation Labels*, chapter 19, pages 291–308. Springer-Verlag, 1997.

[8] K. Silverman, M. Beckman, J. Pitrelli, C. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi : a standard for labeling english prosody. In *ICSLP*, volume 2, pages 867–870, Banff, Alberta, Canada, 1992.

[9] Aubergé V. *Developing a structured lexicon for synthesis of prosody.*, chapter in Talking Machines: Theories, Models, and Designs, pages 307–321. Elsevier Science, 1992.

[10] J. Vaissière. The use of prosodic parameters in automatic speech recognition. In H. Niemann et al., editor, *Recent Advances in Speech Understanding and Dialog Systems*, volume F46. NATO ASI Series, Springer-Verlag Berlin Heidelberg, 1988.

Le même est-il un autre ?

Indices prosodiques de désambiguïsation de certaines répétitions à l'oral.

Christel Leleu

Equipe d'accueil EA 1482 : Recherche sur le Français contemporain.
Université Paris III, 13 rue de Santeuil. Institut de Phonétique, 19 rue des Bernardins 75005 Paris France
e-mail: leleu@cicrp.jussieu.fr

Abstract

This study is a part of a research in progress on prosodic clues to the formulation work in spontaneous French (repetitions, deletions and syntactic breaks). Its aim is the identification of the clues which intervene in the desambiguization of a sequence such as "c'est **très très** au point", the orthographic transcription of which does not show whether it corresponds to an accident of a repeating type (one language object) or to an intensive intention (two language objects). We present the results of two preleminary hearing tests to categorize 10 speech sequences including the repetition "très très". An acoustic analysis, realized independantly, leads to the same conclusions. We thus obtain characteritic prosodic features for each category.

1. Introduction

Dans le cadre de nos recherches sur les indices prosodiques en jeu dans le travail de formulation du français spontané, à situer dans le cadre encore plus large des études menées à Paris III sur la syntaxe de l'oral [Morel & Rialland 92], nous avons recueilli un corpus d'énoncés (extraits d'enregistrements d'oral divers d'une durée totale de 10 heures) contenant des ratures, des ruptures et des répétitions.

Or il s'est avéré que la relecture de la transcription orthographique grossière - c'est-à-dire sans annotation phonético-prosodique - de certaines de ces répétitions a posé problème sur le plan interprétatif : s'agissait-il de répétitions de type "**intensif**" dirons-nous, avec un redoublement de l'objet répété pertinent linguistiquement, ou de répétitions de type "**reprise**" avec un redoublement "accidentel", fruit de la recherche lexicale réelle ou simulée du terme qui suit ? Seul un nouveau recours à l'écoute nous a permis d'en juger.

Dans nos transcriptions, nous nous sommes inspirée de la configuration verticale proposée par le GARS [Blanche-Benveniste & Jeanjean 87] pour représenter les répétitions de type "reprise" : nous écrivons les deux unités répétées l'une en dessous de l'autre, de façon à figurer un paradigme ("piétinement" du même site syntaxique). En revanche, nous avons choisi de laisser les répétitions de type "intensif" dans la linéarité : elles figurent l'une à la suite de l'autre sur l'axe syntagmatique.

Ainsi, dans l'exemple :

CHIR. 315 : *sinon c'est **très très** au point*

rien à l'écrit ne nous permet de trancher entre les deux interprétations :

1. "***très très***" (la locutrice insiste) et
2. "***très*** / ***très***" (un certain travail cognitif retarde la formulation et entraîne la répétition de l'unité).

et la question : le même est-il un autre ? se pose effectivement.

Puisque c'est la dimension orale qui permet de désambiguïser, nous formulons l'hypothèse selon laquelle la prosodie entre en jeu de façon déterminante dans ce phénomène. Reste à identifier le ou les indices (le temps, l'intensité et la fréquence du fondamental) pertinents - et sous quelle forme ? - au point que leur implication dans la répétition entraîne le passage d'une interprétation à l'autre.

Dans cet article, nous rendons compte de la catégorisation, faite par 15 auditeurs, de 10 séquences de parole spontanée contenant la répétition "très très", hors (test d'écoute 1) et en contexte (test d'écoute 2). Les résultats d'identification du test 1 ont été traduits en pourcentage et les occurrences ordonnées des plus au moins caractéristiques pour chacune des catégories "reprise" et "intensif".

De façon indépendante, un travail d'analyse acoustique a été mené en comparant les variations prosodiques des mêmes dix occurrences ; les résultats de ce dernier travail corroborent le classement obtenu par l'analyse perceptive et permettent de dégager trois indices prosodiques impliqués dans la caractérisation des deux catégories.

Ce travail préliminaire sur les répétitions à l'oral suscite par ailleurs des questions plus globales de catégorisation syntaxique et d'articulation langue-parole sur lesquelles nous proposons une première réflexion.

Hormis la linguistique, une telle recherche peut intéresser le domaine de la reconnaissance de la parole. Encombré par les traces orales du travail de formulation, ce dernier pourrait, grâce aux indices fournis, faire le tri entre d'une part le redondant pertinent sémantiquement, et d'autre part le redondant redondant, et donc négligeable sur le plan informationnel (même s'il nous semble que ce type d'information est à exploiter sur le plan structurel ; mais c'est une autre histoire...).

2. Catégorisation perceptive de 10 répétitions

2.1 Procédure expérimentale :

Deux tests d'écoute ont été soumis à 15 auditeurs de langue maternelle française, sans restriction concernant l'âge ou la catégorie sociale.

Ces auditeurs ont eu pour consigne de choisir - en cochant les cases correspondantes sur le formulaire à leur

disposition - entre les étiquettes "Reprise" et "Intensif" justifiées au préalable par quelques explications. En cas d'incertitude, un recours à une troisième case "?" était possible.

Les listes sonores des deux tests ont été réalisés à l'aide des logiciels d'édition du signal "Sounddesigner II" et "Peak" sur Macintosh : le signal original - discussion spontanée entre amies enregistrée sur cassette DAT - a été copié sur ordinateur et toutes les séquences contenant des répétitions sélectionnées et rendues facilement accessibles en tant que "régions" listées. Parmi ces séquences, nous en avons choisies 10 en tenant compte des considérations suivantes :

- Equilibre du nombre des occurrences pour chaque catégorie (4 intensifs, 4 reprises et 2 cas ambigus). Ce choix ayant été fondé sur notre propre écoute, il constitue en soi une pré-catégorisation des occurrences retenues.
- Limitation à la répétition de l'unité "très" produite par la même locutrice afin d'éliminer les sources éventuelles de variations à ce niveau.
- Exclusion des séquences caractérisées par un silence entre les deux unités répétées qui induit une interprétation de type "reprise" systématique.

Ces dix séquences ont ensuite fait l'objet d'une nouvelle segmentation qui a permis d'isoler de leurs contextes les répétitions "très très". Ces dernières constituent la base de la liste sonore du test 1 : elles y ont été dupliquées 5 fois (soit un total de 50 occurrences) et distribuées dans un ordre aléatoire ; un silence de 3 secondes sépare chaque occurrence dont le numéro est annoncé (voix enregistrée de l'expérimentateur), repère indispensable dans cette succession rapide du même.
Le test 1 est d'une durée de 3min20.
Le test 2 a été réalisé selon la même procédure mais à partir des 10 répétitions dans leur contexte ; il est donc plus long : 5min30. Le même ordre de passage des séquences est appliqué.

Le nombre total de fois (n sur 75 puisque chaque occurrence est entendue 5 fois par les 15 auditeurs) où une occurrence est identifiée comme une "reprise", un "intensif" ou un "?" par tous les auditeurs est comptabilisé et traduit en pourcentage pour faciliter la lecture des résultats. Il en ressort un score maximal pour chaque occurrence en fonction duquel nous la classons par ordre décroissant dans une des trois catégories en jeu (cf. histogramme de la figure 1).

2.2 Résultats

En fait, deux groupes - "Intensif" et "Reprise" - se dégagent, de 5 occurrences chacun. Aucune occurrence n'a été majoritairement non identifiée (catégorie "?"). En revanche, nous considérons que l'occurrence 5, classée dans le groupe "Intensif", est ambiguë vu la proximité des scores obtenus (41 % contre 38%, et 20 % de "?"). Elle correspond à un des cas ambigus que nous avions repérés au moment du choix des séquences ; le deuxième, l'occurrence 10, est quant-à-elle identifiée comme une reprise bien qu'étant la moins représentative de cette catégorie, ce qui justifie notre écoute.

On constate que les résultats du test 2 ne viennent pas contredire la catégorisation du test 1. Cela valide l'hypothèse selon laquelle les indices prosodiques désambiguïsateurs sont intrinsèques à la répétition.

Indépendamment du confort d'écoute qu'il procure, le recours au "?" (dont on aurait pu se passer puisque chaque occurrence étant réinjectée 5 fois dans la liste à l'insu de l'auditeur, le doute de ce dernier transparaît facilement lors du décompte de ce choix démultiplié) manifeste si les variations acoustiques du contrepoids au score maximal sont attribuées à la catégorie adverse et peuvent être prises en compte, ou non identifiées et être alors jugées non pertinentes.

3. Indices prosodiques dégagés

Les résultats de l'analyse acoustique confirment les deux groupes de 5 occurrences du volet perceptif. La confrontation des deux analyses a permis de dégager des traits pertinents parmi les paramètres suivants :

3.1 Le temps

Un premier travail sur le rapport des durées globales des deux unités répétées nous a amenée à remettre en question la pertinence de ce paramètre. En revanche, ce travail nous a orientée vers la prise en compte du **rapport des durées vocaliques des deux unités** qui semble quant à lui réellement pertinent : il est de **1 dans le cas de l'intensif**, et de **3 à 4 dans le cas de la reprise**, comme le montre la figure suivante qui met en parallèle les pourcentages d'identification des 10 occurrences avec ce rapport (multiplié par dix pour plus de lisibilité) :

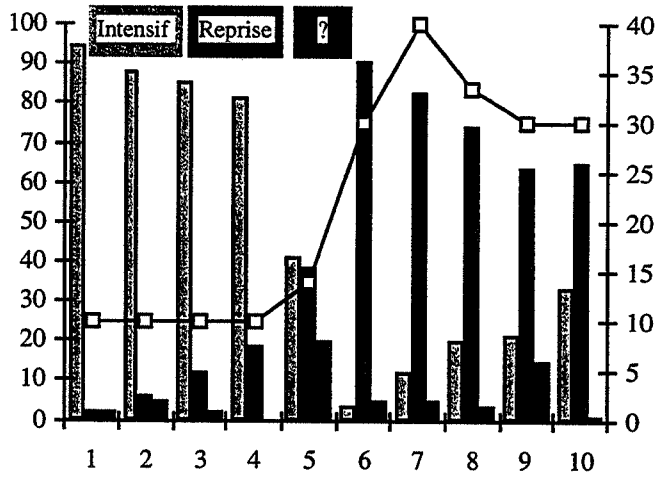


Fig. 1 : Scores perceptifs et rapport des durées vocaliques

Ainsi, pour l'occurrence 4 dont le rapport des durées globales pourrait laisser penser qu'il s'agit d'une reprise, on constate que c'est la partie consonantique qui est allongée mais que les durées des voyelles coïncident (rapport de 1). L'occurrence 5, pour laquelle ce rapport est supérieur à 1 (figure 1), est justement ambiguë.

3.2 Les variations de la fréquence du fondamental

Après observation de la fréquence du fondamental de la deuxième unité répétée, nous sommes arrivée à la

conclusion que cette dernière ne présente pas de variation pertinente pour le phénomène qui nous intéresse. Il n'en est pas de même pour la **fréquence de la première unité** dont le "profil" varie de façon représentative : **plat** sur toute sa longueur **pour la reprise** et **descendant pour l'intensif**. L'occurrence 5, déjà mentionnée pour son ambiguïté, se caractérise notamment par un profil plutôt plat relativement aux autres intensifs.

Si on établit un parallèle entre les profils obtenus (pour plus de clarté, nous avons relevé et comparé le dixième harmonique des spectrogrammes) et le classement issu de l'analyse perceptive, on note que les occurrences dont l'identification a été la moins évidente montre une combinaison de ces deux profils. Il faudra vérifier sur un corpus plus étendu si le sens de la combinaison est pertinent. Nous serions tentée de penser que si l'on représente la forme de la courbe de fréquence à l'aide des trois points successifs : A point initial, B lieu du changement de pente et C point final, c'est la pente BC qui est décisive.

Notons qu'il n'est pas surprenant que la reprise soit marquée par un profil plat si l'on considère que celui-ci manifeste la tension continue des cordes vocales correspondant à la tension liée au travail de recherche lexicale en cause.

3.3 L'intensité

L'intensité en elle-même ne semble pas jouer un rôle de premier plan, mais on note la tendance générale suivante dont il restera à vérifier la portée :

- Pour l'intensif, l'intensité de la deuxième unité est inférieure ou égale à celle de la première, ce que nous schématisons ainsi : O o et o o
- Pour la reprise, c'est l'inverse : o O et o o

3.4 Bilan

Ainsi, il semblerait que la reprise, outre l'intrusion possible d'un silence entre ses deux objets, se caractérise par la fréquence plate du fondamental de la première unité répétée, par un rapport des durées vocaliques compris entre 3 et 4 (la durée de la deuxième voyelle correspond à environ un tiers de celle de la première) et par une intensité - si elle est marquée - plus forte sur la deuxième unité.

Pour l'intensif, semblent pertinents : la descente du fondamental, le rapport vocalique de 1 et une éventuelle marque d'intensité sur la première unité répétée.

4. Point de vue syntaxique

Réfléchissons à la structure syntaxique de 1. "très fort"
très

relativement à 2. "très très fort".

1. Dans ce premier cas, un "très" se substitue à l'autre. Les deux unités ne peuvent constituer ensemble une unité de rang supérieur, elles sont exclusives. La combinatoire (très très) est donc impossible, intenable ; c'est la relation paradigmatique entre les deux unités qui prévaut, relation que le travail de formulation à l'oral fait souvent émerger. Ce que nous traduisons comme suit :

[(très très) fort] -> [(très fort]
très)

2. Pour l'intensif, établissons un parallèle avec la séquence "jolie petite poupée". On peut dégager les constituants syntaxiques suivants :

[jolie (petite poupée)]

Cette segmentation est validée par la substitution de "poupée" à "petite poupée" dans le contexte "jolie --" (procédure de substitution dans un même contexte qui permet de dégager une classe d'équivalence syntaxique)

Nous pensons que cet "agencement" est applicable à notre exemple, "très fort" pouvant être remplacé par "fort" ; ce qui donne :

[très (très fort)]

Les deux "très" s'ajoutent l'un à l'autre, ils appartiennent à deux constituants syntaxiques différents qui se combinent pour former une unité de rang supérieur. On a vu que la force de cohésion de cette unité est reflétée au niveau rythmique (isochronie des unités pour l'intensif).

Si l'on reprend le parallèle avec "jolie petite poupée", on peut supposer que le "très" solidaire de l'adjectif a un statut particulier relativement au "très" extérieur. En effet, on peut dire : "jolie jolie petite poupée" (jolie (jolie (petite poupée))) ou, dans le même registre syntaxique : "beaucoup beaucoup plus fort" (beaucoup (beaucoup (plus fort))), ce qui laisse supposer que c'est l'élément externe qui est répété. Il faudra voir comment cette dichotomie se manifeste sur le plan prosodique.

5. Une tripartition linguistique

Dans notre corpus, nous avons maints objets qui, répétés, présentent le même "profil" prosodique qu'un intensif ou une reprise mais qui ne posent aucun problème d'ambiguïté du fait de leur catégorie syntaxique. Ce sont (liste non exhaustive constituée à partir de notre corpus) :

- d'une part : les prépositions, les déterminants (articles et adjectifs déterminatifs), les pronoms (relatifs et personnels), les conjonctions de subordination et de coordination, et le présentatif "c'est", qui sont toujours interprétés comme des reprises s'ils sont répétés et relèvent du travail de formulation.
- d'autre part : certains adverbes, les adjectifs, les verbes et les noms, qui sont toujours interprétés comme du linguistiquement pertinent si réitérés.

La catégorie à laquelle "très" appartient est donc "à cheval" entre ces deux listes, basculant dans l'une ou dans l'autre en fonction de ses caractéristiques suprasegmentales.

Une question : les répétitions marquées lexicalement le sont-elles aussi prosodiquement ?

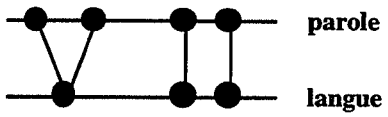
6. Articulation langue-parole

C'est à ce niveau de l'analyse que se pose la question du même et de l'autre. En effet, dans les deux types de répétitions analysées, deux occurrences sont produites ; on a bien deux objets de parole. Mais dans un cas, ils

correspondent à deux objets de langue (intensif), et dans l'autre à un seul (reprise) ; aussi aboutit-on à ces équations à l'impact déjà éprouvé :

2 = 2 (intensif)
2 = 1 (reprise)

Ce que nous représentons provisoirement de la façon suivante (la première tire figurant le codage de la répétition et la seconde son décodage) :



très très très
très

Ainsi, pour en revenir à la question posée en titre, le même au niveau de la parole est un autre dans la mesure où le doublet est linguistiquement pertinent, ce qui est le cas pour l'intensif. En revanche, en ce qui concerne la reprise, le même consiste en quelque sorte à la "réflexion" d'un même objet (parallèle avec l'image) ; une des unités - laquelle ? - est "évacuée" de l'axe syntagmatique dans lequel la linéarité de l'oral le loge, et on ne peut l'appréhender qu'en tant que substitut de l'autre unité.

Nous serions tentée d'appeler **"répétition"** ce dernier phénomène de reprise, et **"réduplication"** celui pour lequel la réitération est pertinente linguistiquement : un morphème, le même (copie d'un schéma de base qui nous rappelle certaines théories phonologiques), est ajouté à un autre dont il "développe" un des traits sémantiques. C'est une première réflexion sur le sujet.

Dans notre corpus, on trouve de telles "réduplications" dans la deuxième liste lexicale présentée en 4. Par exemple :

SANSLOGIS 3320 : *tous ces gens **subissent subissent subissent***
ALG. 5140 : *le **grand grand** problème*

Notons que des études sur la perception des répétitions de mots montrent que ces dernières passent souvent inaperçues lors d'une écoute "normale" dirons-nous, c'est-à-dire non focalisée sur ce phénomène [Clancy 1972]. Ce qui semble aller dans le sens de la langue.

Nous sommes assez séduite par la distinction que [Blanche-Benveniste 97] emprunte à [Olson 94], entre le "dire" et "le vouloir dire", le travail de formulation se situant pensons-nous à ce premier niveau. Elle écrit à ce propos :
"En tant que participant à un dialogue, nous sommes surtout attentifs à ce que l'autre veut dire, plus qu'à ce qu'il dit, et à la forme exacte et littérale de son discours. C'est sans doute ce qui explique que nous sommes finalement peu gênés par les répétitions, les hésitations et les redites propres au langage parlé improvisé, que nous percevons à peine (et qui semblent insupportables quand on les met par écrit)."

Citons également [Boomer et Laver 68] traduits par [Blanche-Benveniste & Jeanjean 87, p. 152] à propos de cette "langue" - ce "vouloir dire" ? - vers laquelle le locuteur et l'interlocuteur tendent, au mépris de la parole dite : "dans la conversation ordinaire, tout se passe comme si nous passions une sorte de contrat social tacite, à la fois en tant que locuteurs et en tant qu'interlocuteurs, pour écarter ces phénomènes de notre claire conscience ; comme si nous regardions, au-delà, l'énoncé régulier et idéalisé".

7. Conclusion et perspectives

Cette étude est donc un travail préliminaire sur les répétitions à l'oral. Elle opère un va et vient entre perception et production qui nous semble indispensable pour identifier les indices prosodiques pertinents dans le travail de formulation.

Les premiers résultats obtenus, à savoir la pertinence du rapport des durées vocaliques, de la forme du fondamental du premier objet répété et de l'emplacement du pic d'intensité, restent à vérifier sur un corpus plus étendu par le biais de la même procédure et de manipulations de type modifications des paramètres prosodiques.

La question de l'existence d'indices prosodiques pertinents pour les unités des deux listes dressées en 4. se pose également.

L'ambiguïté observée ici est un matériau intéressant pour réfléchir à l'articulation langue-parole, aux manifestations des différentes classes lexicales du français relativement au travail de formulation, à la dichotomie que nous avons posée - et qui reste à valider - entre "répétition" et "réduplication", à l'impact des relations syntaxiques entre les unités au niveau prosodique.

Bibliographie

[Blanche-Benveniste & Jeanjean 87] : Blanche-Benveniste C. & Jeanjean C., 1987, Le français parlé. Transcription & édition, Didier-Erudition, Paris.

[Blanche-Benveniste 97] : Blanche-Benveniste C., 1997, Approches de la langue parlée en français, Ophrys.

[Boomer et Laver 68] : Boomer et laver, 1968, "Slips of the tongue", *British Journal of Disorders of communication*, 3, pp. 2-11.

[Clancy 72] : Clancy, P. 1972, "Analysis of a conversation", *Anthropological Linguistics*, 14-3, pp. 78-86.

[Morel & Rialland 92] : "Emboîtements, ruptures, autonomie dans l'intonation française", Travaux du CERLICO, Tome 1 : La subordination (10-11 juin 1991), pp. 221-243.

[Olson 94] : Olson D.R., The World on paper. The conceptual and cognitive implications of writing and reading, Cambridge University Press, Cambridge.

Tous mes remerciements pour leur aide à Christine et Jean-Pierre Leroy, Bernard Gautheron et Lamine Badaoui.

Association prosodie-syntaxe : validation par synthèse

Philippe Martin

Département d'Etudes Françaises, Université de Toronto
Carr Hall, 100 St Joseph St
Toronto, Ontario, Canada M5S 1A1
Tél.: 416 966 2631- Fax: 416 920 4634
e-mail: pmartin@chass.utoronto.ca

ABSTRACT

In phonosyntactic models of intonation, the prosodic structure, encoded by melodic contours located on stressed syllables, is associated to the syntactic structure either strictly, (when there is homomorphy between syntax and intonation), independently (where intonation is considered totally independent from syntax), or conditionally. In the latter case, specific rules and constraints (stress and syntactic clash conditions, maximal number of syllables of a prosodic word, etc.) govern the possible association mechanism that links one or more possible prosodic structures to a given syntactic structure. This paper present experimental result based on the perception of synthesized sentences, in which various prosodic structures are associated with the same text. In order to examine the acceptability of the prosodic variants, contours were realized by combinations of three different parameters, fundamental frequency, duration and pauses.

1. INTRODUCTION

Les modèles phonologiques récents de l'intonation comportent une structure prosodique, généralement encodée par des contours mélodiques liés aux syllabes accentuées de la phrase, qui entretient avec la structure syntaxique une relation d'association gouvernée par des contraintes prosodiques propres. Ni nécessairement homomorphe à la structure syntaxique, ni complètement indépendante, la structure prosodique est associée à la syntaxe dans la mesure où les contraintes de collision d'accent, de collision syntaxique, de taille de mot prosodique et d'eurythmie sont satisfaites [Mar87], [Del95].

Ces contraintes, qui portent sur les mots prosodiques (unités minimales qui contiennent une seule syllabe accentuée et qui sont organisées hiérarchiquement par la structure prosodique) sont :

Collision d'accent : 2 syllabes accentuées successives doivent être séparées par une syllabe inaccentuée ou par une pause, ou encore par une séquence suffisamment longue de consonnes. Ainsi dans

Jean-Luc Hess parle souvent à la radio

l'accent normalement placée sur *Luc*, dernière syllabe de *Jean-Luc*, se trouve décalée à gauche sur *Jean* du fait de la collision avec l'accent de *Hess*. Le maintien de la syllabe accentuée sur Luc demanderait l'insertion d'une pause entre *Jean-Luc* et *Hess*. D'autre part, la présence d'une syllabe accentuée précédant immédiatement Jean aurait alors provoqué la désaccentuation de Jean-Luc, placé à un niveau inférieur dans la structure syntaxique.

De même, dans

La course Vendée-Globe est dangereuse

la syllabe accentuée de Globe entraîne le déplacement de l'accent de Vendée sur la première syllabe (qui est souvent perçue pour les non-initiés comme « *la course Vents des Globes* »). Ici la présence de deux consonnes terminant *course* permet le maintien de la syllabe accentuée sur la première syllabe de *Vendée*.

Collision syntaxique : la structure prosodique ne peut regrouper des mots prosodiques correspondant à des unités syntaxiques de dernier niveau appartenant à des groupes syntaxiques différents [Mar87]. Ainsi, Dans

((L'ami) (d'Antoine)) ((truquait) (les courses))

dont l'organisation syntaxique limitée aux mots prosodiques (unités ne contenant qu'un seul accent tonique) est indiquée par un parenthésage, une structure prosodique inadmissible grouperait *d'Antoine* et *truquait*, qui correspondent à des unités dominées par des nœuds distincts dans la structure prosodique.

**[[L'ami] [[d'Antoine] [truquait]] [les courses]]*

Un mot prosodique contient au maximum 6 (7) syllabes [Del95]. Dans le même exemple, *l'ami d'Antoine truquait les courses*, la désaccentuation d'un ou plusieurs des mots prosodiques *l'ami*, *d'Antoine*, *truquait* et *les courses*, entraîne la formation de mots prosodiques plus grands, qui ne pourront contenir plus de 6 (7) syllabes (la valeur de ce paramètre dépendant de la vitesse d'élocution). Seront donc possibles selon cette contrainte, les regroupements suivants :

[[L'ami] [d'Antoine]] [[truquait] [les courses]]

[[L'ami d'Antoine]] [[truquait] [les courses]]

[[L'ami] [d'Antoine]] [[truquait les courses]]

Chacun de ces regroupements se fait de manière à respecter la condition de non-collision syntaxique. En revanche, le regroupement des 2 premiers mots prosodiques dans

[L'arrière grand-père] [d'Elisabeth] [truquait] [les courses]

n'est pas possible, le mot prosodique *L'arrière grand-père d'Elisabeth* contenant 8 syllabes.

Eurythmie : parmi toutes les structures possibles, satisfaisant à ces contraintes, qui peuvent être associées à une structure syntaxique donnée, la plus probable dans le choix du locuteur sera la plus eurythmique [Wio85], [Mar87]. Dans *L'ami d'Antoine est amusant* le regroupement prosodique en *[l'ami d'Antoine] [est amusant]*, dans lequel les deux mots prosodiques comportent chacun 4 syllabes, est plus eurythmique que *[[[l'ami] [d'Antoine]] [est amusant]]*, dont les mots prosodiques ont successivement 2, 2 et 4 syllabes.

Pour une expérimentation perceptive de ces contraintes, vois aussi [Dom81] et [Gro83].

2. VALIDATION PAR SYNTHESE

L'observation des données ne permet pas toujours de collecter toutes les réalisations possibles des structures prosodiques, et d'établir ainsi que les structures qui ne respectent pas les contraintes rappelées plus haut ne sont effectivement jamais réalisées. La synthèse permet alors de générer des configuration théoriquement non acceptables et le jugement auditif d'établir si des traits prosodiques spécifiques rendent des associations particulières prosodie-syntaxe mal formées. Une étude de ce type avait déjà été décrite [Mar78], mais en utilisant des procédés de synthèse moins convaincants, ce qui pouvait entacher les résultats des tests de perception.

2.1 Procédure expérimentale

Les phrases de synthèse ont été réalisées à l'aide du programme WinPitch [Win97], permettant la modification très rapide et conviviale de phrases naturelles préalablement enregistrées et analysées. Pour restreindre la combinatoire et neutraliser en grande partie les effets de la contrainte d'eurythmie, on a choisi des phrases de 4 mots prosodiques comportant le même nombre de syllabes (2, 3, 4 et 5 syllabes) :

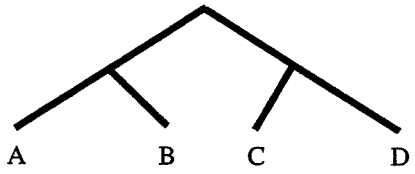
(P2) *L'ami d'Antoine truquait les courses*

(P3) *Les pylônes de l'hydro sont tombés sur la route*

(P4) *Le président du syndicat a rassuré les députés*

(P5) *Les annulations de réservation avaient troublé la majorité*

Ces 4 phrases présentent la même hiérarchie syntaxique au niveau des groupes contenant un seul accent (autre qu'un accent d'insistance ou d'emphase) :



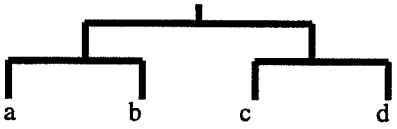
((L'ami) d'Antoine)) ((truquait) (les courses))

((Les pylônes) (de l'hydro)) ((sont tombés) (sur la route))

((Le président) (du syndicat)) ((a rassuré) (les députés))

((Les annulations) (de réservation)) ((avaient troublé) (la majorité))

Pour ces 4 exemples, la structure prosodique homomorphe à la structure syntaxique comporte 4 mots prosodiques a, b, c et d organisés selon la hiérarchie :



Les mots prosodiques sont donc pour **P3**, les tronçons de prosodie (évolution de la fréquence fondamentale, de l'intensité des durées syllabiques) de :

(a) *les pylônes*

(b) *de l'Hydro*

(c) *sont tombés*

(d) *sur la route*

et la séquence de contours prosodiques correspondant à la structure donnée plus haut est :

- Montant	**+ Montant**	**(+ Montant)**	**- Montant**
	+ Ample	**- Ample**	
- Long	**- Long**	**- Long**	**+ Long**

Les traits phonologiques utilisés dans le tableau ci-dessous se manifestent clairement dans la courbe mélodique expérimentale :

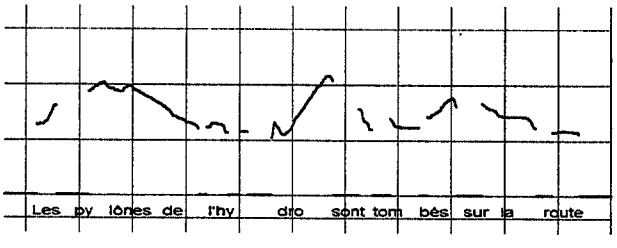


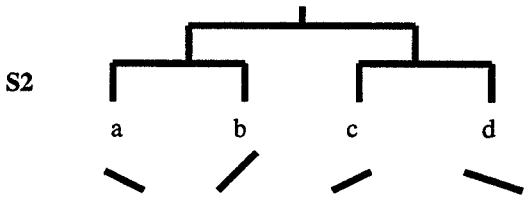
Figure 1: Courbe mélodique de **P3**

Le test de perception a porté essentiellement sur des réalisations présentant des collisions syntaxiques, et encodées par une combinaison des 3 paramètres suivants :

- Accent mélodique : variation de fréquence fondamentale;
- Durées : variation de durée syllabique;
- Pause : présence absence d'une pause ;

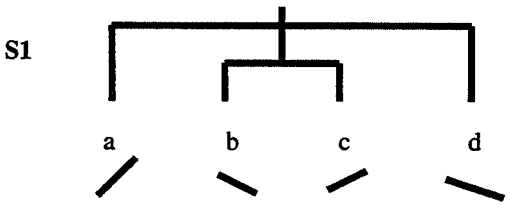
Ces différents paramètres ont été utilisés seuls ou en combinaison, de manière à obtenir 8 possibilités de réalisation d'une structure prosodique (Contour mélodique seul, durées seules, pauses seules, contour + dur, contour + durées + pauses, etc.).

2.2 Collision syntaxique



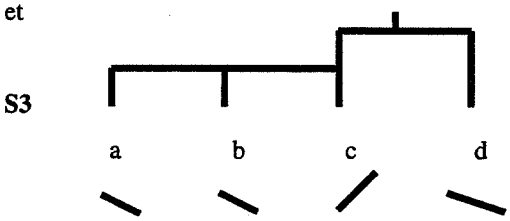
Dans notre exemple à 4 mots prosodiques, la structure prosodique homomorphe à la structure syntaxique instaure une coupure de premier niveau entre les mots b et c, ce qui correspond au parenthésage syntaxique. Une coupure entre a et b ou entre c et d provoque une collision syntaxique en regroupant les unités syntaxiques correspondantes alors qu'elles relèvent de syntagmes distincts. Toutefois la coupure entre l'avant dernière et la dernière unité est une collision admise [Dell84].

Les structures synthétiques associées aux phrases **P2-P5** sont donc :



+ Montant	- Montant	+ Montant	- Montant
+ Ample		- Ample	
- Long	- Long	- Long	+ Long

et



- Montant	- Montant	+ Montant	- Montant
- Long	- Long	- Long	+ Long

2.3 Procédure expérimentale

Dans une première phase, les 4 phrases ont été prononcées par une seul locuteur et modifiées par une synthèse temporelle de type additif de manière à encoder les 2 structures prosodiques testées, référées par les préfixes **S1** et **S3** (**S2** étant le cas d'homomorphie). Les traits utilisés à la synthèse pour l'encodage des différents structures ont combinés la variation mélodique, la durée et la pause.

Les différentes réalisations (29 au total) ont été ensuite soumises à un test auditif portant sur le degré de naturel jugé sur une échelle allant de 1 à 5. Dans la phase préliminaire, onze auditeurs ont noté ces phrases, placées sur un site Internet permettant une procédure moins contraignantes pour les sujets participants.

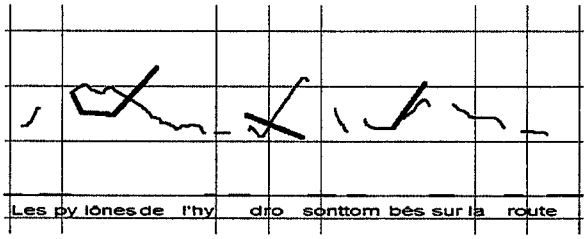


Figure 2: Exemple de synthèse de la phrase P1S3FR

Dans cet exemple, les contours synthétiques apparaissent en trait épais. Les syllabes correspondantes sont respectivement portées à 150 ms, 90 ms et 150 ms de manière à réaliser les contours théoriques nécessaires.

3 RESULTATS

Les résultats bruts des tests de perception sont consignés dans le tableau 1 :

	Moyenne	Ecart type
1 S1P2F	2.5	1.0
2 S1P2FP	2.6	0.8
3 S1P2FRP	2.3	1.0
4 S1P2P	3.5	1.2
5 S1P3F	3.7	0.8
6 S1P3FR	2.9	1.4
7 S1P3FRP	2.5	0.9
8 S1P4F	4.1	1.0
9 S1P4FR	3.6	1.1
10 S1P4FRP	3.4	1.2
11 S1P4P	4.2	0.9
12 S1P5F	3.5	0.8
13 S1P5FR	3.3	0.8

14 S1P5FRP	3.1	0.6
15 S1P5R	3.9	0.9
16 S3P2F	2.7	0.9
17 S3P2FP	2.3	0.5
18 S3P2FRP	2.0	0.7
19 S3P3F	3.1	1.1
20 S3P3FR	2.7	0.8
21 S3P3FRP	2.6	0.9
22 S3P3R	3.2	0.6
23 S3P4FR	4.2	0.7
24 S3P4FRP	4.1	0.7
25 S3P4R	3.8	0.7
26 S3P5F	4.2	0.6
27 S3P5FR	3.8	0.8
28 S3P5FRP	3.5	0.8
29 S3P5R	4.4	0.8

Table 1: Résultats des tests perceptifs.

Dans ce tableau, les 2 premiers caractères se réfèrent au type de structure prosodique réalisée (1 ou 3), les 2 caractères suivants au nombre de syllabes de la phrase (2 à 5), et les caractères suivants correspondent à la présence d'un trait phonétique de fréquence (F), de durée (R) ou de pause (P). Ainsi S3P2FR est une phrase à 3 syllabes par mot prosodique, avec coupure principale entre les 3ème et 4ème mots, et une structure prosodique réalisée par des variations de fréquence fondamentale et de durée syllabique.

3 Interpretations des resultats

L'examen rapide des scores d'acceptabilité, dont la validité statistique ne saurait être pleinement établi vu le nombre de sujets (12) montrent les tendances suivantes :

1. Les structures **S3**, qui sont théoriquement acceptables, sont effectivement mieux acceptées que les structures **S1** qui contiennent une collision syntaxique ;
2. Dans la catégorie **S3**, l'acceptabilité augmente avec le nombre de syllabes du mot prosodique ;
3. Le même effet apparaît dans le groupe **S1**, avec des scores plus faibles ;
4. Lorsque les réalisations sont jugées inacceptables, les marques prosodiques sont synthétisées par la conjugaison d'une variation mélodique et d'une augmentation de durée syllabique correspondante ; Lorsque un seul paramètre, pause, fréquence fondamentale au durée est utilisé, les phrases correspondantes sont perçues comme plus acceptables même en cas de collision syntaxique.

4 Conclusions

Les résultats des test perceptifs sont bien évidemment préliminaires, et ne concernent que 2 structures non homomorphes de phrases à mots prosodiques aux mêmes nombres de syllabes et à la même structure syntaxique. Toutefois, les tendances qui se dégagent des données montrent que l'association d'une structure prosodique provoque une collision syntaxique est effectivement perçue comme peu acceptable, surtout si l'encodage de cette structure par des marques prosodiques se fait par la combinaison d'une variation de hauteur concomitante à un allongement de durée. L'effet est plus marqué si le nombre de syllabes des mots prosodiques est faible, alors que l'effet des pauses ne semble pas déterminant.

Bibliographie

[Del95] Delais-Roussarie, E. (1995), "Pour une structure parallèle de la structure prosodique", Thèse, Université de Toulouse-Le-Mirail.

[Dell84] Dell, F. (1984) "L'accentuation dans les phrases en français", in Forme sonore du langage, Dell, Hirst, Vergnaud réd., pp. 65-122.

[Dom81] Dommergues, J-Y. and Grosjean, F. (1981) "Performance structures in the recall of sentences", Memory & Cognition, Vol 9, 5, pp. 478-486.

[Gro83] Grosjean, F. and Dommergues, J-Y. (1983) "Les structures de performance en psycholinguistique", L'Année Psychologique, Vol 8, 3, pp. 513-536.

[Mar78] Martin, Ph. (1978) "Perception des séquences de contours prosodiques de phrases synthétisées", Actes des IXèmes JEP, Lannion.

[Mar87] Martin, Ph. (1987) "Prosodic and Rhythmic Structures in French", Linguistics, Vol. 25, pp.925-949.

[Ros95] Rossi, M. (1995) "A Principle-based Model for Predicting the Prosody of Speech", in Levels in Speech Communications and Interactions, Sorin et al. Ed., Elsevier, pp. 159-170.

[Win97] WinPitch, logiciel d'analyse temps réel de la fréquence fondamentale, http://www.winpitch.com.

[Wio95] Wioland, F. (1985) "Les structures rythmiques du français", Slatkine-Champion, Paris.

Mesure perceptive de la fonction linguistique de la prosodie pure à l'aide d'énoncés réitérés

Rilliard Albert & Aubergé Véronique

Institut de la Communication ParléeUniversité Stendhal / INPG
B.P. 25 — 38040 Grenoble Cedex 9
email : {rilliard, auberge}@icp.inpg.fr

ABSTRACT

We present here a perceptual measure experiment of the linguistic segmentation hints carried by prosody. We oriented ourselves towards a similitude test between couples of stimuli. The sentences are made of several segmentation variations for group and proposition levels, the couples showing each configuration in pairs. 20 listeners are at the same time able to associate the similar area and syntactic level frontiers. They admit a single syllable forward translation on the reiterated stimuli. They distinguish the couples which do not show the same frontiers. The results also show that the listeners are puzzled (random choices) when the proposed frontiers delimit complex segments.

1. INTRODUCTION

L'une des fonctions bien connues de la prosodie est de délivrer des indices sur la segmentation des énoncés. Ce travail vise à proposer une méthode de mesure de ces informations ; le but idéal, à long terme, étant d'établir une grille de mesures qui, pour chaque fonction de la prosodie (nature de l'énonciation, contexte, locuteur) indique la quantité et la nature des informations linguistiques véhiculées [Hir97]. On obtiendrait en quelque sorte une description de la prosodie naturelle qui pourrait servir de référence à une évaluation de la prosodie synthétique, relativement à un usage donné.

Par ailleurs, nous cherchons à valider les propositions que nous formulons sur un traitement cognitif de la prosodie représenté par une superposition de contours porteurs et portés ancrés en des points de rendez-vous [Aub92]. Chaque niveau de contour correspondant à un niveau linguistique. Le but de cette expérience étant d'aller vers la mesure perceptive de ces points de rendez-vous.

Dans la première approche que nous présentons, très modeste en regards de ces buts, nous nous sommes limités à une prosodie de lecture de phrases isolées, sans consigne contextuelle aux locuteurs. Nous nous plaçons ainsi dans une situation prosodique bien cadrée (ces énoncés sont extraits d'un corpus à partir duquel a été établi un modèle de prosodie synthétique à l'ICP [Aub92]) et de laquelle nous savons pouvoir attendre des informations prosodiques sur la segmentation du même domaine que les informations syntaxiques. À la mesure de la perception quantitative et qualitative des indices prosodiques de segmentation, nous associons une analyse acoustique afin de comprendre quels faits morphologiques (au sens de morphologie prosodique) sont porteurs de ces indices de segmentation.

Mais comment mesurer les effets de segmentation de la prosodie indépendamment des autres « structurateurs » linguistiques ? La solution la plus immédiate est de supprimer de l'énoncé toute information pertinente autre que prosodique, ce qui se produit dès que les informations lexicales ne sont plus accessibles. Visant des buts variés, de nombreux travaux ont abondamment utilisé des procédés de délexicalisation tels que : des glottogrammes [Lho75], des grommelis, des syllabes variées formant un « charabia », ou de la parole réitérée sur une même syllabe canonique ([ma] pour [Lib78] ou [ba] pour [Oll73]).

Notre but étant non seulement de mesurer perceptivement la prosodie pure mais également de mesurer acoustiquement les formes prosodiques discriminantes, nous avons choisi pour notre étude d'utiliser la parole réitérée sur la syllabe [ma]. En effet, l'imitation (*mimicry*) par un auditeur/locuteur d'un énoncé, écouté en boucle perceptive (proposé d'abord par [Lib78]), conserve bien les propriétés prosodiques acoustiques et perceptives du signal de parole [Lar83]. La variation des paramètres prosodiques sera d'autant plus facile à dégager que les valeurs intrinsèques et co-intrinsèques sont en tout lieux identiques et que la nature phonétique des segments permet un étiquetage temporel clair et systématique.

Nous proposons donc une première expérience qui consiste à établir et à vérifier perceptivement le corpus de parole réitérée ; puis une seconde dans laquelle des auditeurs jugent l'association d'énoncés de prosodie pure avec des énoncés complets, écrits et/ou parlés.

2. PROTOCOLE EXPÉRIMENTAL

Nous avons dans un premier temps sélectionné dans le corpus original de notre modèle de synthèse 55 phrases de 4 à 13 syllabes variant par le déplacement des frontières de segmentation Groupe Nominal (GN) / Groupe Verbal (GV) ; Proposition (Prop) Principale (PP) / Prop. Subordonnée (PS) (ou Complétive (PC) ou Relative (PR)) ; ou par l'inversion de ces segments. Une paire minimale du corpus étant opposable soit par un déplacement d'une frontière de même valeur, soit par des

frontières co-occurrentes délimitant des segments de valeur différente.

Deux locuteurs (une femme et un homme) écoutent par trois fois la phrase originale, puis la reproduisent une fois à l'identique (sous sa forme pleine), et une fois sous une forme réitérée en «mamama...». Après acceptation par les locuteurs de leurs productions réitérées, les énoncés originaux sont abandonnés. Les phrases appariées, pleines et réitérées, de chaque locuteur sont présentées en couples à 12 auditeurs qui jugent de la qualité de l'imitation sur une échelle de 0 (différent) à 5 (identique). 22 couples bien notés (note médiane supérieure à 4) sont retenus pour l'expérience suivante.

Le test de dissociation consiste à présenter tous les couples possibles de phrase pleine/réitérée de même longueur à 20 auditeurs selon deux conditions : (i) la phrase pleine est présentée uniquement sur l'écran et la phrase réitérée est présentée à l'oral, et (ii) la phrase pleine est présentée sur l'écran et à l'oral et la phrase réitérée à l'oral.

Une seule écoute est autorisée ; le temps de réponse n'est pas limité. Un choix ternaire entre «*Oui*», «*Non*» ou «*Je ne sais pas*» est proposé en réponse à la consigne : « L'énoncé en «mamama...» que vous entendez est-il compatible avec la phrase pleine que vous venez d'entendre/que vous lisez ? ». Le choix «*Je ne sais pas*» est proposé aux sujets car il s'agit d'une tâche complexe et métalinguistique. On tente ainsi de canaliser une éventuelle classe «poubelle» générée par un choix binaire forcé et d'obtenir des réponses motivées pour les choix « *Oui* » et « *Non* ».

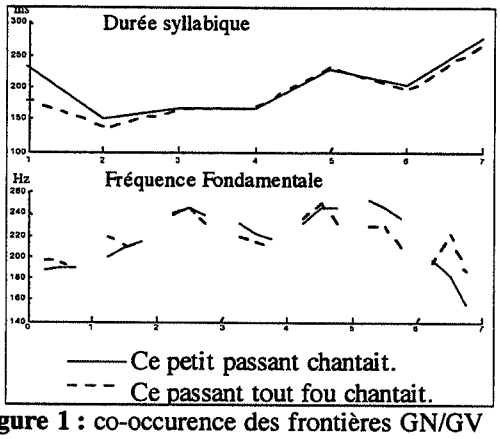


Figure 1 : co-occurence des frontières GN/GV

3. ANALYSE DES RÉSULTATS

On compare les scores d'appariement (réponses «*Oui*» au test) et de dissociation (réponses «*Non*») des couples de phrases. Nous vérifions ainsi que les phrases réitérées sont appariées à leurs énoncés originaux à hauteur de 92% en moyenne, ce qui confirme les résultats de la première expérience.

D'une manière générale, on remarque que les frontières sont d'autant mieux appariées que le contenu des groupes ou propositions est simple : plus le groupe est long, plus les variations sont marquées par les auditeurs. À l'inverse, pour les groupes inférieurs à 3 syllabes, les frontières ne semblent pas perçues. Ceci va tout à fait dans le sens de l'analyse acoustique en terme de contour global et de sa décomposition en contours porteurs et portés [Aub92]. On remarque peu de confusions entre frontières de groupes et de propositions, et un décalage d'une syllabe entre l'emplacement d'occurence des frontières de chaque phrase peut être autorisé (cf. § 3.2.1).

Six phrases contenant deux propositions ont été réalisées avec des pauses silencieuses (réitérées à l'identique). Chaque couple incluant l'une de ces six phrases obtient, comme on pouvait s'y attendre [Gro80] de manière générale, de bons scores d'association ou de dissociation correcte.

Les variations inter-locuteurs sont faibles (corrélation de 0,89 entre les réponses d'associations pour chaque locuteur). D'un autre côté, on observe sur les deux conditions expérimentales, que la présentation orale des deux stimuli a surtout pour effet de diminuer le nombre de réponses « *Je ne sais pas* » des sujets (75 réponses au lieu de 225 – sur 2880 au total), mais ne modifie pas les proportions des résultats en substance (corrélation de 0,69 des réponses « *Oui* » et de 0,64 pour les « *Non* »). Le nombre des réponses « *Je ne sais pas* » est en moyenne inférieur à 10% et culmine à 11% seulement lorsque les sujets associent les stimuli autour de 45% – résultat proche du hasard.

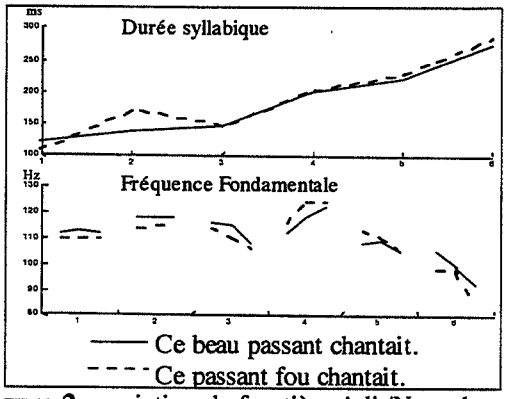


Figure 2 : variation de frontière Adj./Nom dans le GN sujet.

Examinons plus en détail les résultats.

3.1. Co-occurrence des frontières

3.1.1. Frontières de même valeur

Ces frontières sont les mieux associées.

• **Les frontières GN/GV :** ce sont les seules frontières de groupes représentées en nombre suffisant pour une comparaison intra-valeurs. Elles sont associées en moyenne à 87% quels que soient les constituants du GN, même en cas d'inversion Nom/Adjectif à l'intérieur du GN (cf. Figure 1). Cependant les résultats d'association sont d'autant meilleurs que les sous-groupes avec inversion sont petits (moins de 3 syllabes). Nous interprétons ce fait, d'après les principes de modélisation de la

prosodie auxquels nous nous rattachons [Aub92], comme une perturbation dans le décodage des contours porteurs par les indices des contours portés (voir Figure 2).

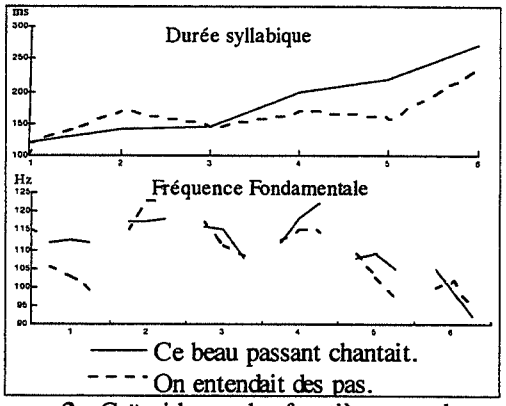


Figure 3 : Coïncidence des frontières sur des segments de valeurs différentes

3.1.3. Frontières de valeurs différentes

• **De même niveau :** les réponses sont proches du hasard : 42% d'association en moyenne pour des frontières coïncidentes mais de valeurs différentes pour un même niveau (groupe ou Prop) (cf. Figure 3). La coïncidence des frontières n'est pas suffisante pour que les auditeurs associent les stimuli, ils récupèrent donc, pour discriminer (47% de dissociation – réponses « Non ») la valeur du groupe ou de la proposition.

Les scores montent à 73% d'associations pour les inversion PP/PS quand la structure en groupes dans chaque proposition coïncide. (e.g. un cas – 2 couples – de Prop/Prop est associé à 73% : *Quand l'enfant pleurait, il était malade.* vs. *Ce passant chantait quand Toto dansait*)

• **De niveau différent :** la dissociation est meilleure (65% vs. 25% d'association) quand les niveaux des segments sont différents (groupe vs. Prop).

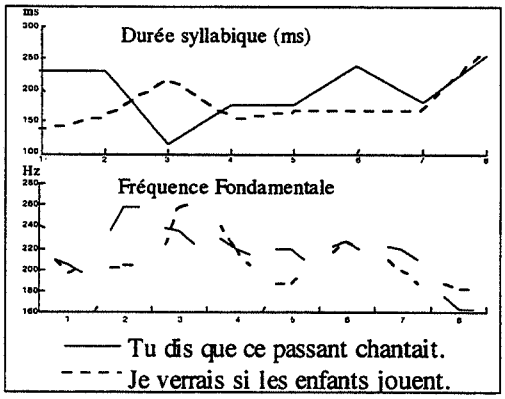


Figure 4 : frontière de proposition décalée d'une syllabe.

3.2. *Frontières non co-occurrentes*

3.2.1. Frontières de même valeur : Le décalage de la frontière pour une même valeur des structures entraîne une dissociation très nette (73%) des phrases dès la **deuxième** syllabe de décalage. Pour une seule syllabe de décalage, on observe un comportement assez complexe : lorsque la frontière du stimulus réitéré est décalée en avant d'une syllabe par rapport à la phrase pleine (écrite ou oralisée), les scores sont de 50 % de dissociations (cf. figure 4), alors qu'un décalage « en avance » d'une syllabe reste comparable aux autres valeurs de décalages, i.e. 66% de dissociation. Cela pourrait rejoindre la notion de frontière floue proposée par Campbell [Cam93], la dissymétrie devant peut-être être imputée à la tâche non symétrique (plein vs. réitéré) de ces comparaisons.

3.2.2. Frontières de valeurs différentes

• **De même niveau :** Les réponses montrent une dissociation nette des phrases (69% de dissociation contre 23% d'association), avec toutefois une différence entre les scores des stimuli construits avec deux groupes GN/GV (dissociés à 76%) et ceux composés de propositions (dissociés à 62%). Cette différence rejoint le résultat du §3.1. où nous avons observé que les frontières des propositions étaient moins prégnantes que les frontières GN/GV.

• **De niveau différent :** Les réponses donnent pour ce cas des dissociations de 70% en moyenne, avec une différence notable entre les frontières séparées par 2 syllabes (82% de dissociations) et celles n'ayant qu'une syllabe de différence (dissociées à 58 % seulement).

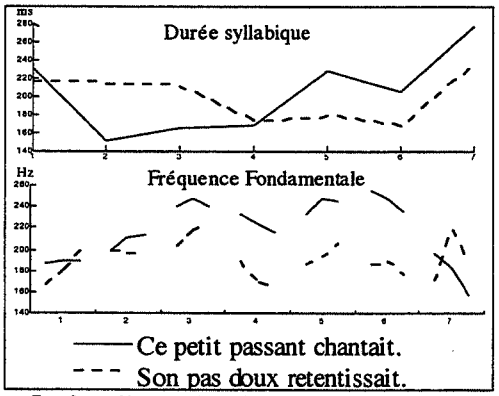


Fig. 5 : frontières séparées de 2 syllabes ; 67% de dissociation.

3.3. *Phrases plus complexes*

Certaines paires de stimuli présentes dans le corpus montrent une structure plus complexe que les cas décrits jusqu'à présent. Nous pouvons cependant essayer de comprendre ce qui se passe alors, en utilisant des combinaisons de cas simples. Nous présenterons rapidement le cas décrit dans la Figure 5. Nous sommes ici en présence d'énoncés pour lesquels les frontières GN/GV ne coïncident pas. Cependant la frontière de l'énoncé réitéré correspond à une frontière syntaxique à l'intérieur du GN de la phrase originale.

Le sujet perçoit donc bien cette marque de frontière, mais ne perçoit pas les autres. Ce cas de figure, pour lequel se manifeste parallèlement une co-

occurrence de frontières, un manque et une différence de la nature des frontières, donne lieu à des baisses de dissociation en comparaison des scores de frontières décalées simples (40% au lieu de 67% de dissociation dans le sens inverse, où le cas est plus simple. Voir Figure 5).

Tableau 1 : valeurs des corrélations entre les phrases réitérées et leurs originaux

		Durée originale	Durée vidée
Loc. 1	corrélation	0,64	0,39
	écart type	0,31	0,32
Loc. 2	corrélation	0,70	0,49
	écart type	0,25	0,22

3.4. Résultats complémentaires

Des calculs de durée et de vitesse d'élocution ont été réalisés pour les syllabes [ma]. Les frontières perçues correspondent à des vitesses ascendantes. Cette adéquation semble encore plus claire si l'on représente la durée des Groupes Inter P-Center (GIPC) [Bar94]. Nous avons également décidé quel meilleur représentant de la durée choisir en « vidant » les suites sonores (pleines et réitérées) de leurs durée intrinsèques et co-intrinsèques pour manipuler des durées syllabiques « pures » (pour une définition, voir [Aub92]). Des trois méthodes, les courbes qui rendent le mieux compte des résultats perceptifs sont celles en GIPC. Les calculs de corrélation entre la durée des phrases réitérées et celle des phrases originales (cf. tableau 1), montrent une corélation plus forte pour les valeurs « brutes ».

4. CONCLUSIONS ET PERSPECTIVES

Cette première expérience de mesure perceptive de la nature et de l'emplacement des informations prosodiques de segmentation est encourageante puisque (i) les auditeurs ne manifestent pas de difficulté à la tâche – qui aurait pu a priori paraître complexe et pour le moins non écologique ; (ii) les auditeurs sont à la fois capables de discriminer les frontières syntaxiques importantes, d'associer les frontières de mêmes valeurs et de dissocier des formes prosodiques véhiculant des segmentations non comparables.

Nous pensons que cette première étape est suffisamment concluante pour pouvoir envisager l'utilisation de ce principe d'association de phrases prosodiques « pures » par réitération, soit par comparaison avec des formes pleines, soit dans des tests de préférence (selon un protocole analogue au test d'évaluation de prosodie synthétique de phrases pleines [Mor98]). Les réponses non exploitables (données au hasard) ne sont pas en contradiction, au contraire, avec les hypothèses que nous supposons par ailleurs être une base cognitive du fonctionnement de la prosodie.

Par contre, il est clair également que le corpus utilisé lors de ce test se révèle pauvre sur plusieurs points : les pauses sont des indices trop forts (il est inutile de comparer des stimuli avec et sans pauses, il est préférable d'établir deux corpus indépendants). Les résultats semblent montrer que les auditeurs reconnaissent les différences de niveau (groupe vs. Prop), il serait intéressant d'enrichir les oppositions à l'intérieur d'un même niveau. Enfin, il serait intéressant de recommencer cette expérience sur des contours porteurs « purs », générés par synthèse et « débarrassés » ainsi de contours portés : on vérifierait ainsi (i) si ce type de stimuli incomplets – impossibles à rencontrer au naturel – permet d'accéder sans bruit à l'information du niveau du porteur (ce qui semble advenir dans [Mor98]) (i) si, en conséquence, le contour porteur est bien un support du processus de décodage perceptif de la prosodie. Des tests de ce type ne peuvent cependant être mis en œuvre sans une connaissance suffisante du mode de transfert des informations prosodiques des phrases originales vers les stimuli réitérés.

BIBLIOGRAPHIE

[Aub92] Aubergé V. (1992), Developping a structured lexicon for synthesis of prosody, in *Talking Machines: Theories Model and Designs*. Bailly G., Benoit C., and T.R. Sawallis (Eds), Elsevier Science, Amsterdam.

[Bar94] Barbosa P. & Bailly G. (1994), Characterization of rhytmic patterns for text-to-speech synthesis. Speech Communication, Vol. 15, pp 127-137.

[Cam93]Campbell N. (1993), Automatic detection of prosodic boundaries in speech. *Speech Communication*, 13, pp 343-354.

[Hir97] Hirshberg J. (1997), Intonation Theory, *ESCA,Workshop on intonation*, Athens, Greece, p19.

[Gro80] Grosjean F. (1980), Linguistic structures and performance structures: Studies in pause distribution. Dans Dechert, H. et Raupach, M. (Eds.). *Temporal variables in Speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 1980

[Lar83] Larkey L.S. (1983), Reiterant speech: an acoustic and perceptual validation. *J.A.S.A.* 73 (4), pp 1337-1345.

[Lho75] Lhote E., Filleau M. & Grange F. (1975), Reconnaissance de patrons intonatifs. *6ème Journées d'Etude de la Parole*, pp 29-37.

[Lib78] Liberman M.Y. & Streeter L.A. (1978), Use of nonsense-syllabe mimicry in the study of prosodic phenomena. *J.A.S.A.* 63 (1), pp 231-233.

[Mor98] Morlec Y., Rilliard A., Bailly G. & Aubergé V.(1998), Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: methodology and first results. *1st Int. Conf. on Language Resources and evaluation*, (to appear).

[Oll73] Oller D. K. (1973), The effect of position in utterance on speech segment duration in English, *J.A.S.A.* 54 (5), p 1235-1247.

L'intonation de la variété francoprovençale parlée à Sarre (Aoste)

Stefania Roullet[1&2] & Antonio Romano[1]

[1] Centre de Dialectologie de l'Université Stendhal - Grenoble III - BP 25 - 38040 Grenoble (France)
Tél.: ++33 4 76 82 43 80 - Fax: ++33 4 76 82 43 56 - e-mail: romano@u-grenoble3.fr
[2] B.R.E.L. - Bureau Régional pour l'Ethnologie et la Linguistique - via Grand Eyvia, 59 - 11100 Aoste (Italie)

Abstract

In this short presentation we would like to resume some observations about the analyses we carried out on the intonational patterns of the francoprovençal variety spoken in Sarre. We propose a model whose aim is to inspect the way statements and interrogations (y/n-questions) are realised by means of different phrase intonation. Furthermore the stylisation criteria here proposed to describe the overall F_0 contours allow us to show some local pitch variations characterising trisyllabic and disyllabic words (paroxytons or oxytons) in different positions within the utterance.

1. Objectifs et méthodologie

Dans le cadre d'un plus vaste travail consacré à l'étude de l'intonation de quelques parlers romans, nous présentons ici les premiers résultats concernant la variété francoprovençale parlée à Sarre (Vallée d'Aoste), obtenus en analysant un corpus de phrases affirmatives (de type déclaratif) et interrogatives (questions totales), avec une approche instrumentale. L'intérêt à l'égard du patois - mais aussi de l'italien régional - que l'on parle à Sarre, représentatif, en quelque sorte, d'une aire linguistique à part entière dans l'espace des variétés d'Italie, est particulièrement important vue l'absence d'études consacrées à l'analyse des aspects prosodiques concernant le domaine francoprovençal, à cheval entre la France, l'Italie et la Suisse [Tua94].

Sur la base d'une approche instrumentale et d'une modélisation empirique, nous avons analysé les trois principaux paramètres acoustiques qui caractérisent les plus importants phénomènes prosodiques, à savoir la durée, l'intensité et la fréquence fondamentale (F_0), ces deux derniers étant considérés dans leur variabilité temporelle. Nous nous sommes intéressés notamment aux valeurs de ces paramètres en correspondance des voyelles, lieu privilégié de la manifestation des variations mélodiques. Nous avons mené ensuite une étude statistique, pour évaluer leur degré de variabilité en fonction du déplacement des unités accentuelles à l'intérieur de la phrase et des différentes façons selon lesquelles s'organisent les groupes accentuels, sous les contraintes imposées par l'intonation de la phrase.

Grâce à un programme que nous avons développé dans un environnement Matlab™, nous avons procédé à l'analyse, au traitement statistique et à la stylisation de l'allure de F_0 de différentes répétitions retenues pour une même phrase, en ayant recours à des critères de segmentation déjà utilisés par d'autres chercheurs [Abry85] et à un algorithme d'extraction semi-automatique de ces paramètres basé sur la méthode de la détection des passages par zéro (*ZC-method*) [Léon70, Boë71].

Nous avons suivi une méthode de phonétique descriptive en essayant d'identifier des modèles intonatifs présentés ici laissant apparaître clairement les éléments qui contribuent à distinguer les deux modalités (affirmative vs. interrogative) du système intonatif de la variété considérée. Ces modèles, reconnus partiellement communs, dans certaines conditions, aux autres variétés romanes parlées par les locuteurs enregistrés (notamment l'italien), ont été évalués par des tests auditifs informels ; ils peuvent cependant faire l'objet d'éventuelles modifications ultérieures, suite à l'analyse, actuellement en cours, de la partie restante du corpus et à leur validation sur des extraits de parlé spontané.

2. Corpus analysé

À partir du moment où il n'existe pas d'études concernant les aspects intonatifs du francoprovençal, à l'exception d'un travail présenté par M. Contini [Cont95] (qui, à titre d'exemple, a voulu prendre en compte les deux modalités d'une phrase prononcée dans le patois d'Aussois, en Savoie), et que, par ailleurs, plusieurs linguistes souhaitent désormais que ceux qui s'occupent des aspects prosodiques aient une connaissance active de la langue qu'ils étudient [Hir], nous avons décidé de privilégier l'analyse des parlers que nous maîtrisons le mieux : ceux de Sarre et de Cogne.

Pour ces deux variétés nous avons constitué un corpus identique, en utilisant un modèle de phrase SN-SV de type SVO, qui est le plus courant et le moins 'marqué'. À partir de ce modèle, nous avons obtenu des énoncés progressivement plus étendus, à l'aide des méthodes d'expansion déjà utilisées par d'autres chercheurs [Cont81, Sor94] et en exploitant toutes les possibilités combinatoires assurées par un ensemble d'éléments lexicaux, représentatifs des types syllabico-accentuels des parlers en question (dans notre cas oxytons et paroxytons, trisyllabiques et dissyllabiques). Le choix lexical à été conditionné par des contraintes d'ordre phonétique, syntaxique et sémantique, qui ont mené à des phrases parfois "inhabituelles", mais acceptées par le locuteur, appelé à les prononcer avec une attitude naturelle.

En effet, notre étude concernant essentiellement les voyelles, nous avons dû nous référer à des contextes consonantiques comparables en retenant des mots constitués d'un maximum d'occlusives sourdes facilitant la segmentation.
Pour ce qui est des expansions utilisées (de simples adjectifs ou des syntagmes prépositionnels comprenant des toponymes), elles ont été associées alternativement aux substantifs du SN et du SV, soit séparément soit ensemble, de façon à obtenir des phrases représentatives du nombre le plus élevé possible de réalisations.
En ce qui concerne la modalité, comme on l'a déjà dit, nous avons concentré notre attention sur les patrons intonatifs les moins connotés, d'un point de vue sémantique, des énoncés affirmatifs et interrogatifs. Pour ces derniers nous avons privilégié la forme qui présente l'inversion du sujet, étant la plus fréquente dans les variétés actuellement en cours d'analyse.
Les interrogatives que nous avons considérées présentent la forme du double marquage [Bla94], une construction qui permet notamment de faire apparaître deux éléments d'un même paradigme sujet ; dans notre cas il s'agit d'une forme lexicale et d'un pronom *conjoint.*
Par exemple, pour la phrase *lo tsareti l a vu lo profète.*, 'le charretier a vu le prophète.', l'interrogative avec inversion correspondante est : *lo tsareti l a ti vu lo profète?*, 'le charretier a-t-il vu le prophète ?'.

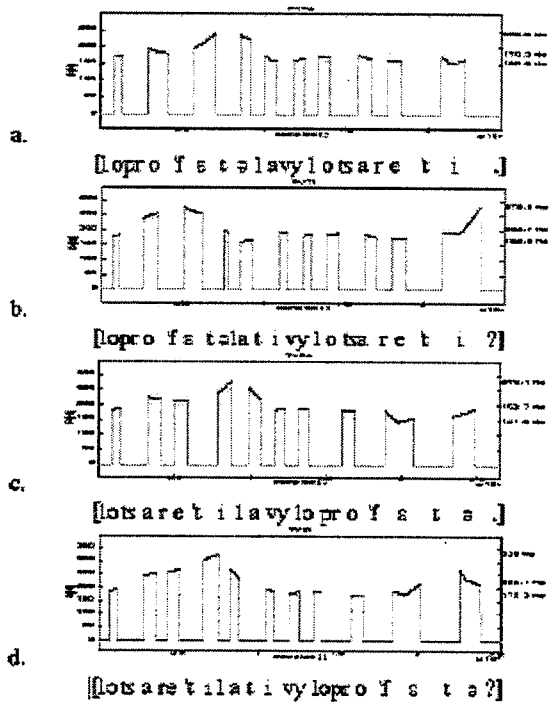


Figure 1. Exemples des patrons intonatifs de base dans la variété francoprovençale de Sarre (locutrice MM49 ; a. et c. affirmatives ; b. et d. interrogatives). Les phrases analysées sont : a./b. 'le prophète a/a-t-il vu le charretier./?' ; c./d. 'le charretier a/a-t-il vu le prophète./?'. Les contours sont représentés seulement en correspondance des voyelles des énoncés. Les contours moyens (plus foncés) sont superposés aux contours des différentes répétitions de chaque phrase, après normalisation temporelle.

Toutefois, puisque la construction avec le double marquage, à cause de l'adjonction d'une syllabe, empêche de comparer de façon satisfaisante les deux modalités, nous avons jugé indispensable d'utiliser également des interrogatives qui ne présentent pas d'inversion, bien que cette forme soit sans doutes moins fréquente.
Pour un seul locuteur, les phrases obtenues à partir d'un même corpus sont : 68 affirmatives et 68 interrogatives sans inversion, chacune répétée 5 fois ; 68 interrogatives avec inversion, chacune répétée 3 fois ; au total : 884 répétitions.
Nous ne présentons que les résultats de l'analyse des phrases simples sans expansions et de celles qui n'en ont qu'une (de type adjectival ou prépositionnel), prononcées par l'un des deux locuteurs retenus pour la variété de Sarre : la locutrice M.M. âgée de 49 ans au moment de l'enquête.

3. Présentation des résultats obtenus

En l'état actuel de nos recherches, le nombre de tracés analysés du corpus, bien qu'encore réduit, nous permet déjà de faire quelques considérations intéressantes. Pour des raisons de manque d'espace, nous ne pouvons présenter que quelques exemples jugés significatifs et qui nous semblent suffisants pour éclaircir notre méthodologie.
En figure 1, nous présentons quatre exemples de courbes intonatives analysées, issues d'un traitement statistique opéré sur la durée et sur les trois valeurs (initiale, finale et intermédiaire, relative à un point jugé significatif) retenues pour F_0, de chaque voyelle, pour toutes les répétitions d'une même phrase.
Sur la base de ces contours moyens, construits selon une méthode s'inspirant de celle déjà utilisée par l'*Institut de Phonétique de Grenoble* (cf. par exemple [Cont81]), nous avons également pu obtenir des fichiers de texte dans lesquels sont indiquées les valeurs numériques moyennes correspondant à la durée (en ms), à une mesure de l'énérgie du signal (en dB), et aux trois valeurs de F0 retenues pour chaque voyelle, ainsi que les valeurs des écarts-type.
L'affichage des courbes, réalisé comme en figure 1, permet de connaître aussi les valeurs des niveaux le plus bas et le plus haut atteints par F_0, tout comme la valeur moyenne de cette variable au cours de l'énoncé.

4. Modèles des contours identifiés

Selon un schéma déjà utilisé pour la comparaison entre deux variétés d'italien régional [Rom97], nous avons élaboré des modèles qui résument les caractéristiques des contours intonatifs, par dessus la structure des unités accentuelles, et qui permettent d'identifier les traits les plus significatifs pour chacune des modalités de la variété analysée.

4.1. Allures de la fréquence fondamentale

Nous ne pouvons pas présenter ici en détail les résultats dont nous disposons déjà : nous nous bornerons par conséquent à en donner une description concise.
Comme on peut le remarquer aisément en observant la figure 1, le contour intonatif du syntagme nominal sujet (SN1) des phrases affirmatives est constitué d'un seul pic, dont la partie finale, descendante, se trouve en correspondance de la dernière syllabe de SN1 ou de la première du syntagme verbal (SV), selon que SN1 se termine par un paroxyton ou par un oxyton. S'il y a une expansion adjectivale, deux pics apparaissent ; le deuxième est nettement plus haut que le premier.
Le SV de la phrase affirmative se caractérise par la présence d'un ou de deux pics (selon qu'il y ait ou non une expansion) assez bas, suivis d'une descente importante qui s'achève approximativement au niveau de la fréquence d'attaque. La dernière voyelle tonique coincide presque toujours avec le minimum de F_0.
Dans la phrase interrogative sans expansions, le contour de SN1 monte au début et ensuite descend de façon abrupte. Si, en revanche, le substantif est suivi d'une expansion, on relève une situation inversée par rapport à celle décrite pour les affirmatives : un pic plus haut est suivi d'un pic moins important.
Malgré la présence de légères oscillations (les pics ne se situent pas toujours sur les voyelles toniques), la première partie du SV des phrases interrogatives, ne se différencie pas trop de celle des affirmatives.
En revanche, sur le dernier mot, les variations de F_0 se font très importantes. En particulier, la courbe présente une forte descente suivie d'une montée sur la tonique qui atteint un niveau très haut et se termine, mais seulement s'il s'agit d'un paroxyton, par une descente très abrupte.
Le schéma intonatif des interrogatives sans inversion ne présente pas de différences remarquables par rapport à celui des interrogatives avec inversion, malgré le nombre différent de syllabes.
Deux éléments importants (résumés dans les schémas en figure 2) semblent donc contribuer à la discrimination de la modalité : de légères différences dans le SN et, surtout, dans la dernière partie de SV, en correspondance du dernier mot prosodique de la phrase. Cela confirme une fois de plus que dans la discrimination de la modalité on ne doit pas considérer seulement la toute dernière partie de l'énoncé, mais également d'autres points-clef, localisés souvent à l'intérieur du SN. À ce propos, il faut ajouter une remarque interessante concernant la différence de registre entre les deux modalités, les phrases interrogatives étant toujours prononcées avec une F_0 moyenne plus élevée (cf. figure 1).
En ce qui concerne les contours intonatifs qui caractérisent localement les unités accentuelles en position finale de SN1 et de SV, faute d'espace pour une description détaillée, nous nous limitons à présenter une reproduction schématique (v. figure 3).

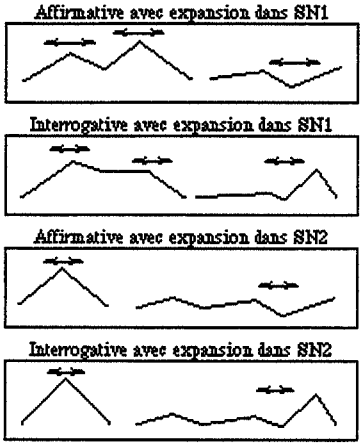


Figure 2. Modèles de contour de modalité des phrases affirmatives et interrogatives (avec ou sans expansion) pour la variété francoprovençale de Sarre. Les flèches indiquent la possibilité pour les pics de se déplacer en fonction de la position de la syllabe structurellement tonique (dans le cas des paroxytons ou des oxytons).

4.2. Réalisation de l'accent

Les données dont nous disposons ne sont pas suffisantes pour établir clairement les paramètres qui entrent en jeu dans la réalisation de l'accent et leur importance hiérarchique.
Sur la base d'une première analyse il semble possible de reconnaître un rôle important aux variations de F_0 et de durée, ce qui est bien connu. En effet, des sommets de F_0 distinguent normalement les voyelles accentuées, à l'exception de la dernière tonique, où se localise un accent de phrase, qui se caractérise par un contour brisé, dans lequel F_0 atteint souvent le niveau le plus bas, et par une longueur considérable.
Il reste encore à vérifier l'importance de l'intensité, en sachant que son développement, globalement descendant, peut être ramené à un comportement défini par des contraintes physiologiques.

4.3. Rôle de la durée

Les données que nous avons obtenues concernant la durée des voyelles confirment les résultats déjà observés pour d'autres langues et résumés dernièrement pour quelques variétés d'italien régional [Sor94], à savoir que les voyelles en position finale de phrase (qu'elles soient toniques ou atones) sont toujours plus longues qu'à l'intérieur de la phrase.
Cet effet de ralentissement final (*prepausal lengthening effect*) semble alors ne concerner que marginalement les voyelles prétoniques du dernier mot du SV, mais intéresser plutôt la tonique et la post-tonique, qui présentent un allongement important par rapport aux

syllabes précédentes. Cela pourrait confirmer l'hypothèse selon laquelle l'allongement final est conditionné par le fait que les syllabes finales véhiculent aussi des informations d'ordre syntaxique permettant la caractérisation de la modalité.
En particulier, on relève que la durée est nettement plus importante si le mot est oxytonique et surtout si la phrase est interrogative. En revanche, si la phrase se termine par un paroxyton, l'allongement se distribue de façon plus uniforme sur les deux dernières syllabes de l'énoncé.

5. Conclusions

En ce qui concerne les allures de F_0, il nous semble intéressant de souligner que les phrases nucléaires analysées pour la modalité affirmative semblent s'articuler globalement sur des variations mélodiques *déclinantes*, avec des sommets plus évidents localisés de préférence au début de la production. Les interrogatives avec un paroxyton à la frontière conclusive de la phrase présentent une clôture 'marquée' (par rapport à celle de la plupart des langues [Hir]), caractérisée par un contour montant-descendant.
Les intonations de base explorées ont révélé combien la structuration intonative semble se baser sur la superposition de contours globaux et de contours locaux, caractérisés par la réalisation d'un accent de phrase, d'accents de groupe et d'accents lexicaux.
Notre interprétation de ces phénomènes se base, dans les grandes lignes, sur le modèle proposé par H. Fujisaki (cf. par ex. [Fuj88]), selon lequel des gestes locaux sont absorbés par des gestes articulatoires globaux qui permettent de traduire les intentions communicatives du locuteur dans des contours réalisées sous des contraintes d'ordre physiologique.

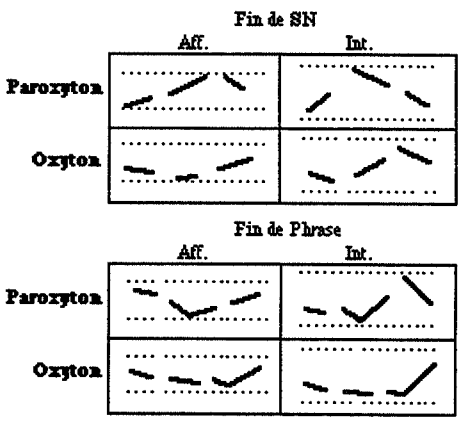


Figure 3. Modèles abstraits des contours des unités accentuelles dans la variété francoprovençale de Sarre quand celles-ci reçoivent un accent de groupe et/ou de phrase (dans SN1 ou bien dans SV).

Ainsi, si d'un côté l'évolution des principales variables de l'organisation prosodique ressent ces contraintes et s'insère dans le contexte des groupes de respiration (cf. [Lie67]), de l'autre elle rend compte localement des commandes linguistiques *d'accent* réalisées pour la caractérisation prosodique de chaque unité accentuelle. Nous pensons que les résultats que nous venons d'exposer sur ces bases, bien qu'encore incomplets, permettent déjà de saisir quelques-unes des caractéristiques prosodiques majeures de cette langue souvent méconnue : le francoprovençal.

Bibliographie

[Abry85] Abry Chr., Benoit Chr., Boë L.J. & Sock R. (1985). Un choix d'événement pour l'organisation temporelle du signal. *Actes des 14èmes JEP*, Paris.

[Bla94] Blasco M. (1994). Les énoncés à "redoublement" et "dislocation" en français contemporain. Analyse en double marquage. *L'information grammaticale*, 63, Paris.

[Boë71] Boë L.J. & Rakotofiringa H. (1971). Exigences, réalisation et limite d'un appareillage destiné à l'étude de l'intensité et de la hauteur d'un signal acoustique. *Revue d'Acoustique*, 4.

[Cont95] Contini M. (1995). Transcrire les faits prosodiques. *Actes de la Conf. ann. sur l'act. sc. du Centre d'Etudes Francopr.*, Aoste, Musumeci.

[Cont81] Contini M., Bellet G. & Boë L.J. (1981). Variations intralocuteurs de F_0 dans la réalisation d'une phrase énonciative en français : normalisation temporelle. *Actes des 12èmes JEP*, GALF, Montréal.

[Fuj88] Fujisaki H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. *Vocal Physiology: Voice Production, Mechanisms and Functions* (O. Fujimura ed.), Raven Press, New York.

[Hir] Hirst D. & Di Cristo A. (à paraître). *A survey of intonation systems*. À paraître.

[Léon70] Léon P.R. & Martin Ph. (1970). *Prolégomènes à l'étude des structures intonatives*. Studia Phonetica, Didier, Paris.

[Lie67] Lieberman Ph. (1967). *Intonation, Perception, and Language*. MIT, Cambridge Mass.

[Rom97] Romano A. & Roullet S. (1997). Analisi intonativa comparata di due varietà di italiano regionale (salentino meridionale e valdostano centrale) sulla base di un corpus fisso di frasi affermative e interrogative, ottenute con espansioni progressive dei sintagmi nominale e verbale. *Actes des VIII Giornate di Studio del G.F.S.*, Pise, 18-20 déc. 1997, (à paraître).

[Sor94] Sorianello P. (1994). Il processo dell'allungamento prepausale: dati ed interpretazioni. *Quad. del Dip. di Ling. dell'Univ. di Firenze*, 5, Florence.

[Tua94] Tuaillon G. (1994). *Le francoprovençal. Progrès d'une définition*. Duc, Aoste.

Caractérisation du débit de parole en français

Brigitte Zellner

LAIP, IMM, Faculté des Lettres
Université de Lausanne, 1015 Lausanne, Suisse
Brigitte.Zellner.@imm.unil.ch

ABSTRACT

This paper is concerned with the evaluation of speech rate in French. Usually, this dynamic parameter is described as a quantitative dimension. It is shown that the slowing down of speech has also major qualitative effects that have to be taken into account. The theory on slowing down speech is thus revised.

1. INTRODUCTION

Généralement, les spécialistes de la parole évaluent le débit de parole d'un locuteur, d'un point de vue strictement *quantitatif*. La vitesse de parole est exprimée en nombre d'unités de parole — segments, syllabes, mots — produites par unités de temps — seconde ou minute — (par exemple, [Gros75]; [Ber96]).

Or, la littérature abonde d'études phonétiques qui montrent que les variations du débit de parole ont des incidences à tous les niveaux de la structure temporelle de la parole. Ainsi, pour le français, ces incidences ont été en particulier relevées au niveau du groupe prosodique (par exemple [Fou95]; au niveau de la syllabe (par exemple [Due87]; [Lac90]; au niveau segmental et articulatoire [Lac90]; [Vax95]. Une caractérisation de la vitesse de parole à l'aide d'un unique paramètre quantitatif n'est donc pas satisfaisante. Cette observation est renforcée par la vérification en resynthèse de la parole. En effet, il ne suffit pas d'augmenter ou de diminuer le nombre de syllabes par seconde pour obtenir d'un synthétiseur, une parole lente ou rapide qui soit proche du naturel.

Ce papier présente un résumé partiel de thèse qui montre que la dimension qualitative doit aussi être prise en compte dans la caractérisation du débit de parole en français. La confrontation de ces résultats avec des études antérieures suggérera que les locuteurs disposent de plusieurs "stratégies" du ralentissement de la parole.

2. METHODOLOGIE

50 phrases déclaratives ont été lues par un locuteur disposant d'une excellente fluidité verbale, d'une part à un débit assez rapide, et d'autre part à un débit ralenti (9 phrases ne dépassant pas 12 syllabes, 20 phrases entre 13 et 20 syllabes, 21 phrases entre 21 et 30 syllabes). Jugées par au moins dix auditeurs natifs du français, ces lectures ont été déclarées hautement intelligibles, sans accent dialectal. Après numérisation, les signaux de parole ont été soigneusement étiquetés par deux experts entraînés pour cette tâche, afin de déterminer la durée de chaque segment. Une seconde analyse a ensuite permis d'isoler les syllabes.

En effet, pour cette étude, l'unité linguistique qui a été retenue est la syllabe, dans la mesure où un consensus assez général s'est établi quant à la réalité psycho-rythmique de cette unité de parole en français (pour un avis différent, voir [Bar94]). Pour ce faire, une grande attention a été portée au problème de la syllabation phonétique. Par exemple, dans le cas d'un mot comme "déguster", il est difficile de prédire ce que le locuteur va produire exactement. Il est possible de syllaber ce mot ainsi: "dé/gu/ster/" ou "dé/gus/ter/" ou encore "dé/gus/ster/", ces trois formes étant tout à fait acceptable en français standard. Egalement les cas d'enchainements et de liaisons ont été vérifiés. Dans tous ces cas, une analyse minutieuse du signal[1] a pu résoudre le problème.

Les durées syllabiques brutes ainsi obtenues ont ensuite été normalisées par transformation logarithmique, afin d'améliorer leur approximation par une distribution gaussienne. Le repérage des frontières prosodiques mineures et majeures a été faite sur la base d'une analyse textuelle proximale, qui examine la succession des mots grammaticaux et des mots lexicaux [Zel96], algorithme proche de celui développé par [Sor87].

Une base de données a ensuite été construite, indiquant pour chaque syllabe le contexte d'occurrence du segment et de la syllabe; la longueur de la phrase, la longueur du groupe prosodique, la nature du mot dans lequel la syllabe apparaît (lexical ou grammatical), la longueur du mot en nombre de syllabes, la position de la syllabe dans le mot, le type de syllabe (description de la complexité de la structure), la durée de chaque composante syllabique (attaque, noyau, rime), la longueur de la syllabe (en nombre de segments), la durée de chaque segment, et la présence éventuelle du schwa. Puis, les 968 syllabes produites en débit rapide ont été comparées aux 1001 syllabes produites en débit lent.

3. RESULTATS

La comparaison entre les deux débits montre, comme attendu, une augmentation du temps de parole (temps de phonation + pauses) en débit lent. Cette augmentation s'explique au regard de différents paramètres dont l'importance est évaluée par rapport à ce temps différentiel – ou ralentissement total calculé pour les 50 phrases.

[1] Cette procédure d'analyse est actuellement testée par une équipe de phonéticiens canadiens, avant d'être automatisée.

3. 1. Mécanisme de ralentissement traditionnellement attendu: la pause

Traditionnellement, il est rapporté dans la littérature que l'un des mécanismes principal du ralentissement du débit consiste en la production de pauses silencieuses (Par exemple, [Sai77]; [Gro75]; [Gro79]; [Bar94). Dans cette analyse, il s'avère pourtant que les pauses produites en débit lent par ce locuteur contribuent seulement pour 7.9% au ralentissement total.

Ce mécanisme se caractérise comme suit dans cette étude. Il est constaté tout d'abord un allongement des pauses déjà existantes en débit rapide. Cet allongement contribue pour 1.07% du ralentissement total.

Le ralentissement est aussi obtenu par production de pauses supplémentaires (6.83% du ralentissement total). En débit ralenti, les pauses émergent dans des groupes initialement longs — et qui ont pu éventuellement être rallongés par la production d'une syllabe supplémentaire. Ainsi, la longueur moyenne des groupes temporels qui sont ainsi ralentis par une insertion de pause est de 7.11 syllabes, différence significative par rapport aux autres groupes en débit lent (F-ratio = 16.91, $p \leq 0.0001$).

Le nombre de syllabes produites entre deux pauses est variable. Les pauses s'insèrent soit à une frontière mineure ou majeure pré-existante, soit à une nouvelle frontière interlexicale — donc souvent après un ralentissement de parole. D'un débit à l'autre, le placement des pauses diffère puisque 57% des pauses sont placées sur une frontière prosodique mineure en débit rapide, contre 71% en débit lent. En débit rapide, les pauses ont tendance à se concentrer non loin de la frontière majeure qui est proche du milieu de la phrase. En débit lent, elles se distribuent sur davantage "d'espace".

Il n'apparaît pas de corrélation entre le débit moyen par phrase et le nombre de pauses. Autrement dit, chez ce locuteur, la vitesse de parole moyenne n'est pas corrélée à la fréquence d'occurrence des pauses.

En conclusion, si le profil de production des pauses se différencie bien entre les deux débits, l'importance des pauses pour ralentir la parole s'avère être un mécanisme beaucoup moins considérable pour ce locuteur, que ne l'aurait laissé suggérer la littérature ([Bar91], [Bar94]). De plus, Barbosa [Bar94]) a suggeré que l'absence ou la présence de pause silencieuse s'inscrit dans le même phénomène d'allongement avec un mécanisme de bifurcation dans la répartition des durées à l'intérieur de l'unité rythmique. Dans cette étude, les évidences empiriques confirment que les pauses et les allongements sont des indicateurs de la structuration temporelle d'un énoncé, un ralentissement de parole signalant une moins grande proximité interlexicale. Par contre, il n'apparaît pas d'évidence que ces deux mécanismes de ralentissement seraient "équivalents" en termes de stratégie phonatoire, dans la mesure où l'un introduit un arrêt et l'autre une continuité du geste verbal. Les distributions de durées des rimes devant pause et des autres rimes sont sensiblement les mêmes dans les deux débits, suggérant qu'il n'y ait pas une différence aussi fondamentale entre les deux groupes de rimes, particulièrement en débit rapide (figure 1).

3.2. Mécanisme moins attendu: la réalisation d'un plus grand nombre d'unités sonores

L'analyse de nos données montre qu'en débit lent, 50 syllabes supplémentaires ont été réalisées pour lire les 50 phrases, contribuant pour 8.8% au ralentissement total. Chez ce locuteur, ce mécanisme de ralentissement est donc légèrement *plus puissant* que celui de la production de pauses silencieuses.

80% de ces syllabes se situent à l'intérieur d'un groupe prosodique et dans 88% des cas, il s'agit d'un cas de réalisation de schwa, 10% des autres cas étaient dues à des diérèses (Ex: in/te/llec/tuel -> in/te/llec/tu/el), et 2% des cas d'hyperarticulation. (Ex: a été (ai)guillé -> a été **ai**guillé).

Dans les cas de réalisation du schwa, la syllabe qui précède a typiquement une durée proche de la durée syllabique moyenne, tandis que la durée de la syllabe avec schwa est plus courte que la durée syllabique moyenne. 86% de ces syllabes se situent en finale de lexème. Enfin, 3% des ces réalisations de schwa concernent des mots grammaticaux (en particulier les déterminants et les pronoms).

Le changement du débit affecte donc la chaîne parlée du point de vue du nombre de syllabes produites et des frontières de syllabes. Dans cette étude, la réalisation d'un plus grand nombre de syllabes est un mécanisme aussi puissant que celui des pauses.

3.3. Un mécanisme important: l'allongement des segments

Le moyen le plus puissant pour ralentir la parole consiste en l'allongement de la durée segmentale comme le montre la comparaison des durées moyennes standardisées. Ce mécanisme de ralentissement contribue pour 69.96% du ralentissement total. Ainsi, du débit rapide au débit lent, la durée d'un segment augmente *en moyenne* de 34.5%. Mais cette moyenne cache des réalités très diverses. Par exemple, plus spécifiquement, dans ce corpus, le segment [l] est allongé de 24%, le segment [˙] est allongé de 2%, le segment [e] est allongé de 63%. Le regroupement en classes de durées, en fonction de la durée moyenne du segment, (cf. table 1) permet d'observer que certains segments n'appartiennent pas à la même classe de durée selon le débit. Par exemple le segment [˙] qui appartient à la classe {˙, Â, d} en débit rapide appartient à la classe {˙, j} en débit lent. Ceci provient probablement d'une part du fait des différentes propriétés d'élasticité des segments — il est plus facile de prolonger un [Â] qu'un [˙] — et d'autre part, de leur rôle phonologique — par exemple, le schwa qui a un statut particulier en français, à une durée assez stable, quel que soit le débit.

Ainsi, un mécanisme puissant du ralentissement de la parole consiste à allonger la durée des unités de parole.

3.4. Effets sur la cohésion interlexicale des énoncés

Un des effets du ralentissement de la parole est qu'il modifie la structure de cohésion interlexicale par l'insertion de syllabes supplémentaires et de pauses. Du fait que le mécanisme des pauses est particulièrement actif aux frontières prosodiques, les enchaînements sont plus concernés que les liaisons.

En effet, dans cette étude, la répartition des liaisons et enchaînements dans les énoncés analysés montrent que pour les deux débits, ces deux phénomènes se répartissent différemment. Les tests du Chi 2 (Rapide: $\chi 2 = 21.778$, $p= 0.0001$, $df=1$; Lent : $\chi 2 = 19.036$, $p= 0.0001$, $df=1$) indiquent que dans les deux débits, l'hypothèse d'une répartition uniforme des liaisons et enchaînements doit être rejetée. Il apparaît un phénomène d'attraction des liaisons à *l'intérieur* du groupe prosodique tandis que les enchaînements sont peu sensibles aux frontières, i.e., ils se distribuent également à l'intérieur des groupes et aux frontières des groupes.

L'effet de syllabes supplémentaires et d'insertion de pauses est donc particulièrement sensible au regard des enchaînements de syllabes. Le locuteur a produit 18% d'enchaînements de plus en débit rapide qu'en débit lent. Par contre, le nombre de liaisons est resté stable malgré la variation de la vitesse de parole. 63% des enchaînements supprimés sont remplacés par l'insertion d'une syllabe avec schwa suivi d'une pause.

Dans cette étude, le processus d'enchaînement est gêné en débit lent par la structure phonologique de la chaîne parlée. Le plus grand nombre de schwas en finale de lexèmes et l'insertion de pauses freinent ce processus de liage interlexical.

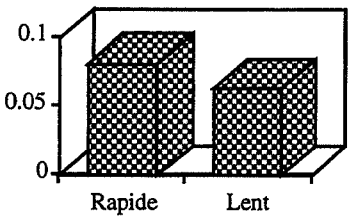
4. DISCUSSION

La théorie du ralentissement la plus souvent rapportée dans la littérature suggère que le parole est ralentie principalement de deux manières: par insertion de pauses et par allongement des unités sonores (par exemple, [Bar94]). Généralement, l'importance de ces deux mécanismes n'est pas évaluée, et elle a souvent conduit à des surestimations de ces mécanismes dans les algorithmes utilisés en synthèse de la parole.

Dans cette étude, la comparaison systématique de 50 phrases lues dans deux débits chez un locuteur montre que le ralentissement peut s'obtenir, par ordre d'importance, par l'allongement des segments, par la production d'un plus grand nombre de syllabes et par l'insertion de pauses. Au niveau qualitatif, un effet important de ce ralentissement est qu'il altère les liens interlexicaux. La prise en compte de cet effet qualitatif permettrait sans doute de mieux caractériser le débit d'un locuteur.

Ainsi, un premier indicateur de proximité entre les mots peut être exprimé par le rapport entre le nombre d'enchaînements et le nombre de mots[2] (qui ne change pas d'un débit à un autre). Plus cet indicateur de proximité est élevé, plus le degré de cohésion interlexicale est élevé. Pour cette étude, on obtient le profil suivant:

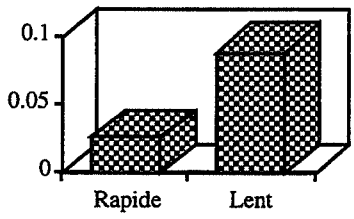
Nb enchaînements / Nb mots



Cependant, pour avoir une bonne appréciation du niveau de cohésion entre les mots, il faut aussi tenir compte du phénomène pausal. En effet, il a été montré au point 3.4 que les pauses ont tendance à freiner le liage interlexical.

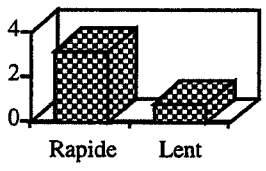
Un indicateur de *distance* interlexicale peut être exprimé par le rapport entre le nombre de pauses et le nombre de mots. Plus cet indicateur est élevé, plus la distance interlexicale est élevée. Pour cette étude, on obtient le profil suivant:

Nb pauses / Nb mots



Un indice de cohésion interlexicale correspond alors au rapport entre ces deux indicateurs de proximité et de distance. Il permet d'exprimer pour un débit donné, le degré de proximité entre les mots en prenant en compte les enchaînements et les pauses. Plus l'indice obtenu est élevé, plus le degré d'intégration temporelle est élevé. Pour cette étude, on obtient le profil suivant:

Indice de cohésion interlexicale



Ainsi, l'indice de cohésion interlexicale est plus élevé en débit rapide du fait du plus petit nombre de pauses et du plus grand nombre d'enchaînements.

L'intérêt de cet indice est qu'il permet de capter un phénomène bidimensionnel, caractéristique du changement de débit en français. Lors de l'évaluation du débit de parole, outre les paramètres quantitatifs, il sera sans doute utile de vérifier le niveau de l'indice de cohésion interlexicale.

[2] Le mot est ici compris au sens où un traitement de texte décompte les unités graphiques.

CONCLUSION

Cette étude montre que l'évaluation du débit de parole devrait prendre en compte des paramètres quantitatifs et qualitatifs.

En français, hormis [Lac91] qui a surtout analysé la dimension phonologique du phénomène, le mécanisme de ralentissement de la parole par production de syllabes supplémentaires n'avait pas été explicitement relevé et évalué dans la littérature. Chez notre locuteur, ce mécanisme est aussi puissant que celui de la production de pauses. Il apparaît aussi que l'émergence des pauses n'est pas liée, ici, à l'allongement des unités sonores précédentes, contrairement aux résultats de [Bar94]. Ces différences dans les résultats avec ceux des autres chercheurs pourraient s'interpréter dans le sens d'une multiplicité des stratégies du ralentissement de la parole.

Table 1: Répartition des segments en classes de durées (rapide, n=2175, lent, n=2295)

Classes du débit Rapide	Durée (log)	
1	1.572	j,w, l, ?
2	1.659	˙, Â, d
3	1.750	n, y, ë, v, i, u, J, O, g, E, z
4	1.842	e, a, b, m, t
5	1.939	E$, O$, ^, Ø, A$, o, k, π, p, π$, s, f
6	2.038	S

Classes du débit Lent	Durée (log)	
1		
2	1.668	˙, j, ?
3	1.753	l, ë, w, Â, d, y
4	1.831	n, v, i, g, z, O
5	1.935	J, b, u, E, m, p, t, a, k
6	2.055	e, A$, o, f
7	2.125	O$, s, π, E$, S, ^, Ø

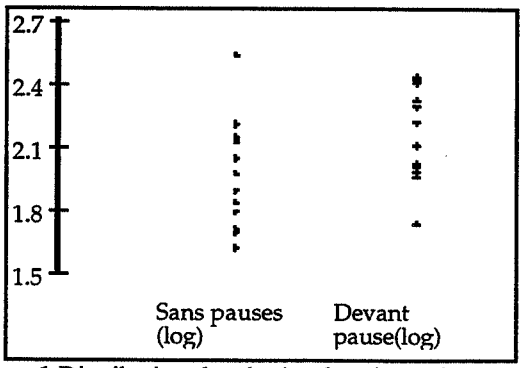


Figure 1 Distribution des durées des rimes devant pause et des autres rimes (valeurs logs) en débit rapide

BIBLIOGRAPHIE

Bar94]) Barbosa, P. A. (1994). *Caractérisation et génération automatique de la structuration rythmique du français.* Thèse de Doctorat. U.R.A. CNRS n°368 - INPG/ENSERG, Université Stendhal, Grenoble.

[Bar91], Bartkova, K. (1991). Speaking rate in French application to speech synthesis. *XIIème Congrès International des Sciences Phonétiques,* (pp. 482-485). Aix en Provence. Actes.

[Ber96] Bertrand, R., et Casolari, F. (1996). Approche prosodique et pragmatique des modulations. *XXIes Journées d'Etudes sur la Parole,* (pp. 179 - 182.). Avignon.

[Due87] Duez, D. , & Nishinuma, Y. (1987). Vitesse d'élocution et durée des syllabes et de leurs constituants en français parlé. *Travaux de l'Institut de Phonétique d'Aix, 11.* 157-180.

[Fou95] Fougeron, C. & Jun, S-A. (1995). Properties of french intonation at fast speech rate. *XIIIème Congrès International des Sciences Phonétiques, 3* (pp. 488-491). Stockholm.

[Gro75] Grosjean, F. , & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica, 31.* 144-184.

[Gro79] Grosjean, F. , & Collins, M. (1979). Breathing, pausing, and reading. *Phonetica, 36.* 98-114.

[Lac90] Lacheret–Dujour, A. (1990). *Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur.* Thèse de doctorat de l'Université de Paris 7.

[Lac91] Lacheret–Dujour, A. (1991). Le débit de la parole: un filtre utilisé pour la génération des variantes de prononciation en français parisien. *XIIème Congrès International des Sciences Phonétiques* (pp. 194-197). Aix en Provence. Actes.

[Sai77] Saint-Bonnet, M., Boe, J. (1977). Les pauses et les groupes rythmiques: leur durée et disribution en fonction de la vitesse d'élocution. *8èmes Journées d'Etude sur la Parole,* (pp. 337- 343). Aix en Provence.

[Sor87] Sorin, C., Larreur, D. & Llorca, R. (1987). A rythm-based prosodic parser for text-to-speech systems in French. *XIème Congrès International des Sciences Phonétiques* (pp. 125-128). Talinn, Estonie. Proceedings.

[Vax95] Vaxelaire, B. (1994). Variation de geste et débit. Contribution à une base de données sur la production de la parole, mesures cinéradiographiques, groupes consonantiques en français. *Travaux de l'Institut de Phonétique de Strasbourg, 24.* 109-146.

[Zel96] Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée.* 1. (pp.7-23).Paris.

REMERCIEMENTS

Mes remerciements à Gerard Bailly (ICP) et Eric Keller (LAIP) pour leurs précieuses suggestions.

Synthèse

Optimisation d'un modèle prosodique pour la synthèse par règles à partir du texte en français.

Albert Di Cristo, Philippe Di Cristo et Jean Véronis

Laboratoire Parole et Langage, Institut de Phonétique d'Aix

29, avenue Robert Schuman, 13621, Aix-en-Provence Cedex 1, France
Tel: +33 4 42 95 36 19 - Fax +33 4 4259 50 96
e-mail: albert.dicristo@lpl.univ-aix.fr

ABSTRACT

This paper presents the theoretical background and the implementation rules of a prosodic model which is integrated into a French text-to-speech system developed in Aix (SYNTAIX). According to this model, prosody is generated, via a set of projections devices, from a phonological representation of rhythm and intonation. This approach allow us to obtain high quality results at a very minimal implementation cost. It enables also to capture a realistic modelling of some aspects of the prosodic variability observed in connected speech.

1. Introduction.

La collaboration entre les Equipes "T.A.L.N". et "Prosodie" du Laboratoire Parole et Langage de L'Institut de Phonétique d'Aix a conduit à la réalisation de deux systèmes complets de synthèse à partir du texte qui sont à présent accessibles sur le réseau Internet[1]. Ces deux systèmes se distinguent par le fait que la génération de la prosodie procède d'une approche stochastique pour le premier [Vér 97] et d'une approche paramétrique à base de connaissances linguistiques pour le second [Dic 97]. Etant donné que les deux approches utilisent le même système de codage prosodique, en l'occurrence, l'alphabet INTSINT [Hir 98], la première approche peut être considérée, indépendamment de sa finalité propre, comme une heuristique permettant d'éprouver la validité des options du modèle linguistique impliqué dans la seconde. Ce qui dénote en fait la complémentarité des deux systèmes.

Nous présentons dans cette communication les fondements théoriques et les phases de mise en oeuvre du modèle prosodique qui constitue le moteur de la seconde approche. L'application de ce modèle est subordonnée à trois conditions.

• La première concerne son adaptation aux performances actuelles du T.A.L.N. qui ne permettent pas de réaliser avec un taux d'erreur acceptable une analyse syntaxique approfondie du texte ou de sa structure informationnelle.

• La seconde condition fait référence à la flexibilité du modèle qui doit être capable d'évoluer en fonction des progrès réalisés dans le traitement automatique du texte. C'est la raison pour laquelle nous défendons l'idée d'un système prosodique noyau, indépendant du style, mais capable cependant d'en simuler la variation par le jeu de ses paramètres.

• Enfin, la troisième condition se rapporte à l'optimisation du modèle qui vise à produire une synthèse de qualité pour un coût d'implémentation minimal. Il n'est pas rare en effet qu'un système performant utilise plus d'une centaine de règles prosodiques. Nous voudrions montrer que l'on peut obtenir des résultats comparables avec un nombre de règles inférieur à dix. Cette économie drastique tient selon nous, d'une part au pouvoir de généralisation du modèle phonologique qui constitue l'entrée de la composante prosodique et, d'autre part, aux dispositifs de projections qui sont mis en œuvre pour générer le signal acoustique à partir d'une représentation abstraite de la prosodie.

2. Architecture du système et fondements théoriques du modèle prosodique.

Notre système de synthèse comporte deux modules: un module linguistique et un module phonético-acoustique. Le module linguistique est constitué de deux sous modules: un module T.A.L.N. et un module prosodique phonologique, ce dernier comportant trois composantes: métrique, tonale et temporelle.. Faute de place, nous n'exposerons pas ici le contenu du premier sous-module, qui a été développé à l'aide des outils du projet MULTEXT [Vér 94]. On précisera seulement que la méthode de "parsing grammatical" (PG) que nous utilisons est adaptée de l'algorithme "chink'n chunk" de [Lib 92]. Elle permet cependant, à la différence de ce dernier, d'effectuer les groupements des items lexicaux compris entre deux signes de ponctuation selon une procédure non-déterministe qui introduit une variabilité structurelle attestée en français parlé.

Le postulat de base du modèle énonce que la prosodie du texte est dérivable (via le PG) à partir des représentations abstraites du rythme et de l'intonation. Ces représentations, qui recourent au formalisme de la grille métrique étiquetée [Hal 95] et au codage de l'intonation en termes de segments tonals [Pier 95], prennent appui sur le modèle phonologique de la prosodie du français

développé par [Hir 84], [Dic 93] et [Dic97].

Le second postulat du modèle stipule que la représentation de l'intonation présuppose celle de l'organisation métrique. L'élaboration de la grille métrique, qui constitue donc la phase initiale du sous-module prosodique, est fondée sur la mise en œuvre de principes et de paramètres généraux, ainsi que de principes et de règles propres au français. Parmi les premiers, le *Principe de Projection* [Ids 92] est décisif, car il établit un lien rigide entre l'assignation d'une tête (ou proéminence) et la formation d'un constituant, ce qui permet de rendre compte conjointement des deux aspects majeurs de la structure prosodique que sont l'accentuation et la constituance. En ce qui concerne le français, la construction de la grille est principalement fondée sur l'application de deux principes:

- *Le Principe de Bipolarisation Accentuelle* [Dic 98] qui attribue un accent initial au premier mot accentogène et un accent final au dernier mot accentogène des groupements issus du "parsing grammatical". Nous obtiendrons ainsi, par exemple, les schèmes accentuels suivants: la DEcouVERT(e) de L'AUTR(e), la DEcouverte de L'AUTR(e), la REuNION de L'AssemBLEE, la REunion de l'AssemBLEE, etc. (les syllabes accentuées sont indiquées en majuscules).

- *Le Principe de Subordination Accentuelle*, qui spécifie les niveaux relatifs de proéminence selon l'ordre suivant: Accent Nucléaire (Unité Intonative)> Accent Final (Primaire)> Accent Initial (Secondaire). Ce principe permet de gérer la hiérarchie des proéminences dans le cadre du mot et au-delà de ses limites.

Le Principe de Bipolarisation Accentuelle fait partie d'une stratégie de "parsing prosodique" qui est formulée par les quatre règles de base décrites dans [Hir 84] (formation des Unités Intonative, proéminence initiale, proéminence finale et réorganisation des proéminences en fonction de leur proximité). Ce principe permet d'intégrer l'accent initial dans une double perspective: rythmique et emphatique. Il rend également compte de la stratégie de "l'arc accentuel" dont l'usage fréquent est attesté dans différents styles de parole [Fon 80] et dans diverses formes de focalisations [Dic 98].

3. Mise en œuvre du modèle

3.1 Le module phonologique

La première étape de mise en œuvre du modèle concerne la construction de la grille métrique à partir du parsing grammatical et, partant, la spécification des niveaux de proéminence (figure 1). L'organisation tonale du texte est dérivée en deux temps selon des procédures de traitement lié qui projettent d'une part la structure tonale profonde à partir de la grille métrique et, d'autre part, la structure tonale de surface à partir de la représentation de cette structure profonde. La première opération revient en fait à associer des gabarits tonals aux constituants métriques définis par la grille (figure 2).

- L-L et L-H, pour les constituants métriques de premier ordre (ou Unités Intonatives), qui correspondent à des segments de texte compris entre deux signes de ponctuation;

- L-H pour les constituants métriques de second ordre (ou Unités Tonales), qui sont équivalents aux groupes accentuels.

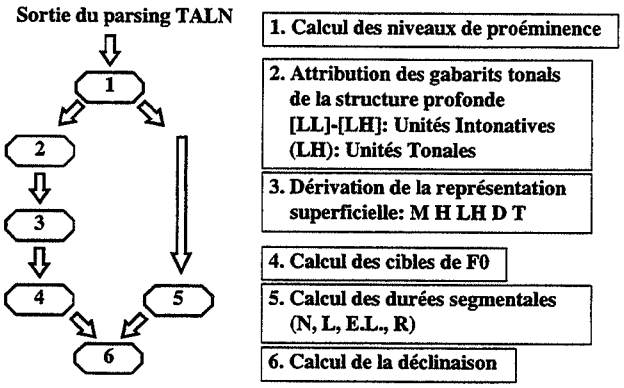


Figure 1: Etapes du module prosodique.

Cette représentation hiérarchique non-linéaire est soumise à la contrainte universelle de linéarisation qui projette les segments tonal sur une ligne unique que nous considérons comme la représentation phonologique superficielle de l'intonation du texte [Hir 98]. La représentation superficielle découle de l'application d'un ensemble de principes et de règles, notamment de la règle d'abaissement (ou de downstep) qui s'applique soit localement, comme marque de frontière de l'UI [Dic 96], soit itérativement, dans le cas d'une question, par exemple [Dic 93]. La représentation superficielle est encodée au moyen de l'alphabet INTSINT [Hir 98]. La mise en œuvre du module phonologique métrique et tonal implique les règles 1, 2 et 3 de la figure 1.

Le modèle intègre également une représentation de l'organisation temporelle, qui est fondée sur une quantification des durées segmentales en quatre catégories, ces dernières étant associées (selon la même procédure de traitement lié) à la structure métrique des énoncés:

- *Normal* (o): forme non-marquée (syllabes métriquement faibles)

- *Lengthened* (+): qui s'applique aux segments des syllabes finales des groupes accentuels mineurs délimités par un accent primaire.

- *Extra-Lengthened* (++): qui s'applique aux segments des syllabes finales des groupes accentuels majeurs délimités par un accent nucléaire.

- *Reduced* (-): qui s'applique aux segments inclus dans un "arc accentuel", afin de simuler les variations du tempo solidaires de ce type de groupement (figure 2).

Cette quantification, qui se fonde sur des données acoustiques et perceptives [Ast 95], [Land 97], peut s'appliquer à chaque segment constitutif de la syllabe: attaque, rime, coda, pour rendre compte des catégories accentuelles proposées par [Dic 98].

3.2 Le module phonético-acoustique.

Ce module comprend les trois étapes correspondant aux règles 4, 5 et 6 de la figure1.

- En premier lieu, une synchronisation est effectuée entre les segments tonals de la représentation superficielle et la rime des syllabes auxquelles ils sont rattachés. Dans la version actuelle du système, deux solutions sont retenues pour l'alignement des tons H:

• en début de rime pour un H lié à une proéminence initiale d'arc accentuel;

• au centre de la rime pour un H rattaché à une proéminence d'Unité Tonale (groupe accentuel mineur).

Les tons de frontière, T ou B, correspondant respectivement à la frontière non-terminale et terminale des Unités Intonatives, sont alignés avec la fin de la rime. L'abaissement local (ou downstep) est positionné au centre de la rime de la pénultième syllabe de l'UI.

Entre deux tons H consécutifs, l'alignement du ton L est moins contraignant, comme semblent le montrer les tests perceptifs que nous avons réalisés. Cette flexibilité peut ainsi être exploitée pour simuler une certaine variabilité.

- Les valeurs cibles de F0 associées aux segments tonals de la représentation phonologique de surface, sont calculées par référence aux données perceptives de [Ros 72]. Cette procédure permet de fixer les registres du locuteur et de calibrer sa dynamique tonale.

- Les durées segmentales sont définies à partir de la durée moyenne de chaque phonème (déterminée sur la base du corpus EUROM 1 [Chan 95]), à laquelle est appliquée un facteur d'élasticité *k* correspondant au niveau de quantification temporelle concerné (cf. supra: N, L, E.L, R). Ce facteur s'applique soit globalement, soit individuellement, aux éléments constitutifs de la syllabe: attaque, noyau et coda [Ast 95].

- Les "effets d'abaissement" globaux sont implémentés en recourant à trois catégories de gammes tonales:

- Une gamme *Gi* correspondant à l'attaque de l'UI;

- Une gamme Gf (réduite par rapport à Gi), correspondant à la fin d'une UI non finale de paragraphe.

- Une gamme Gp (réduite par rapport à Gf) pour les UI finales de paragraphe.

La ligne de déclinaison proprement dite est définie au moyen d'une fonction asymptotique ou d'une simple interpolation linéaire entre les tons représentatifs des registres du locuteur (M, T, B).

En dernier lieu, le signal acoustique est généré à l'aide du synthétiseur par diphones MBROLA développé à Mons [Dut93].

4. Conclusions

La mise en œuvre du modèle prosodique présenté dans cette communication permet d'obtenir une synthèse performante, qui se distingue notamment par sa "fluidité". La flexibilité du modèle rend possible l'adoption de plusieurs stratégies d'encodage rythmique, qui actualisent en fait le caractère probabilitaire de l'accentuation du français et qui concourent ainsi au renforcement du naturel. De plus, le modèle est évolutif, dans la mesure où les principes fondamentaux qui en constituent la base pourront s'appliquer sans difficulté au marquage des diverses formes de focalisation lorsque cette information sera disponible. En l'état actuel, il mise surtout sur la simplicité. Dans cette optique, il permet d'obtenir une synthèse robuste, qui ne nécessite pas une analyse syntaxique approfondie du texte et dont le coût d'implémentation s'avère extrêmement réduit.

BIBLIOGRAPHIE

[Vér97] Véronis, J., Di Cristo, P. & Courtois, F. (1997), "A stochastic model of intonation for French text-to-speech synthesis", Eurospeech'97, vol 5., pp.2643-2346.

[Dic97] Di Cristo, A., Di Cristo, P. & Véronis, J. (1997), "A metrical model of rhythm and intonation for French text-to-speech synthesis", Esca Workshop on Intonation (Athens), pp. 83-86.

[Hir98] Hirst, D.J., Di Cristo, A. &, Espesser, R. (1998), "Levels of representation and levels of analysis for the description of intonation systems", Prosody: Theory and Experiments, (Horne, M.,ed.), Kluwer, in press.

[Vér94] Véronis, J., Hirst, D., Espesser, R. & Ide, N. (1994), "NL and speech in the Multext project", AAAI'94 Workshop on Integration of Natural Language and Speech, pp.72-78.

[Lib92] Liberman, M. & Church, K. (1992), "Text analysis and word pronunciation in text-to-speech synthesis", Advances in Speech Signal Processing, pp. 791-831.

[Hal95] Halle, M. & Idsardi, W. (1995), "General properties of stress and metrical structure", Handbook of Phonological Theory, pp. 403-443

[Pier80] Pierrehumbert, J. (1980), The Phonology and Phonetics of English Intonation, MIT Press.

[Hir84] Hirst, D.J. & Di Cristo, A. (1984), "French intonation: a parametric approach", Die Neueren Sprachen, 83, pp. 554-569.

[Dic93] Di Cristo, A. & Hirst, D.J. (1993), "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français", T.I.P.A., 15, pp. 9-24.

[Dic97] Di Cristo, A. & Hirst, D.J. (1997), "L'accentuation du français: stratégies et paramètres", Polyphonie pour I. Fónagy, l'Harmattan, pp. 77-101.

[Ids92] Idsardi, W..J. (1992), The Computation of Prosody, Ph. D. Dissert. MIT.

[Dic98] Di Cristo, A. (1998), "Vers une modélisation de l'accentuation du français", soumis à Journal of French Language Studies.

[Fón80] Fónagy, I. (1980) "L'accent en français: accent probabilitaire", L'accent en Français Contemporain, pp.123-233.

[Dic96] Di Cristo, A. & Hirst, D.J. (1996), "Vers une typologie des Unités Intonatives en français", Actes des XXIes JEP (Avignon), 219-222.

[Dic93] Di Cristo, A. & Hirst, D.J. (1993b), "Prosodic regularities in the surface structure of french questions", Working Papers , 41, pp.268-271

[Ast95] Astesano, C., Di Cristo, A. & Hirst, D. (1995), "Discourse-based empirical evidence for a multi-class stress system in French", Proc.XIIth I.C.Ph.S., pp. 630-633.

[Ros72] Rossi, M. & Chafcouloff, M. (1972), "Les niveaux intonatifs", TIPA, 1, pp.167-176

[Chan 95] Chan, D. et al. (1995), "EUROM - A spoken Language Resource for the EU.", Proceedings of ESCA, Eurospeech'95

[Dut93] Dutoit, T & Leich, H. (1993), "MBR-PSOLA: Text-to-speech synthesis based on an MBE resynthesis of the segments database", Speech Communication, 13, 435-440.

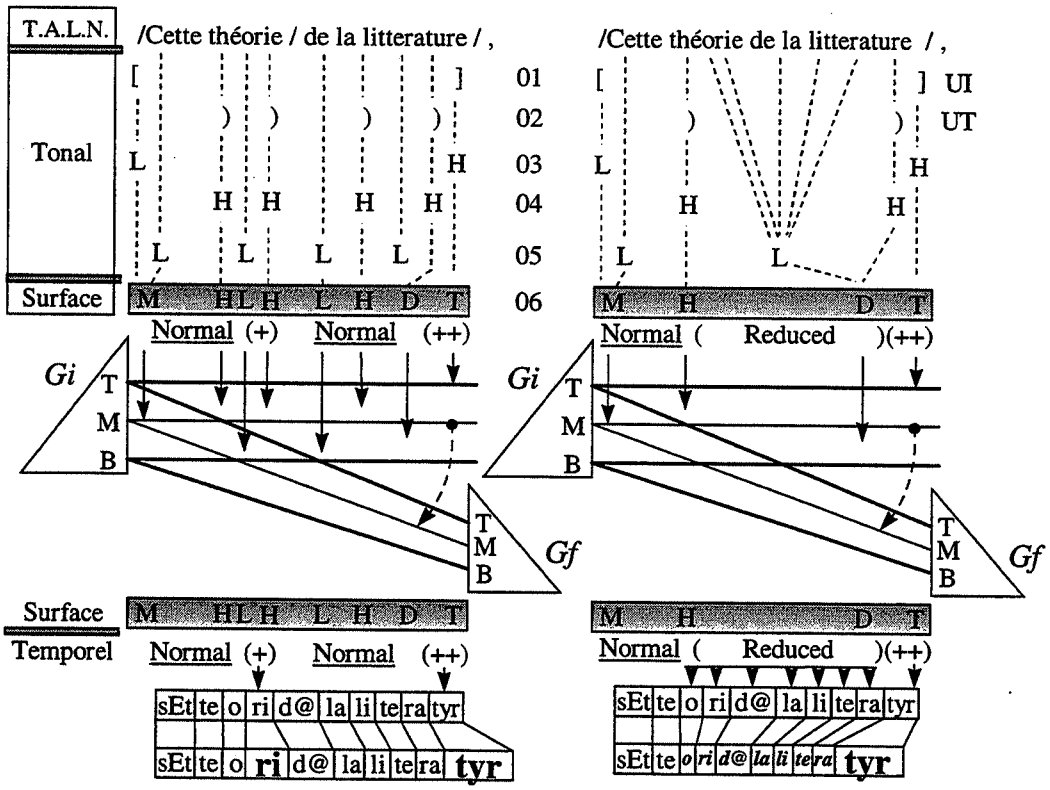


Figure 2: Illustration des dispositifs de projections mis en œuvre dans la génération de la prosodie

[1] *Modèle stochastique* : http://www.lpl.univ-aix.fr/cgi-bin/stolpl.cgi

Modèle paramétrique : http://www.lpl.univ-aix.fr/cgi-bin/metlpl.cgi

Le système de synthèse FIPSVox : syntaxe, phonétisation et prosodie.

Arnaud Gaudinat et Jean-Philippe Goldman

LATL – Département de linguistique
Université de Genève
CH-1211 Genève - Suisse
Tél.: ++41 22 705 72 74 - Fax: ++41 22 705 79 31
e-mail: {gaudinat,goldman}@latl.unige.ch - http://latl.unige.ch

ABSTRACT

This paper describes the text-to-speech system for the french developped at LATL and its applications. FIPSVox is based on FIPS, a large-scale, multi-purpose, GB-based syntactic parser which produces detailed analyses. This syntactic information is directly exploited by 1. the grapheme-to-phoneme module to handle denasalisation, heterophone homographs, French elision and liaison phenomena, and 2. the prosody generation module to determine the dependancy between phrases and identify special syntactic structures such as extrapositions of a constituant (as in cleft, heavy-NP shift and left-dislocation structures), as well as some cases of parentheticals to derive of appropriate prosodic patterns.

1. INTRODUCTION

Un système de synthèse de la parole à partir du texte a pour but de produire un signal acoustique ressemblant le plus possible à une prononciation humaine. La représentation classique d'un tel système comprend plusieurs modules: une conversion graphèmes-phonèmes (ou phonétisation), qui consiste à transformer le texte en une suite de phonèmes ; une génération des paramètres prosodiques qui détermine un profil mélodique idoine permettant d'obtenir un signal acoustique le plus naturel possible ; durant la dernière ét0ape, le synthétiseur génère le signal sonore en fonction des éléments déterminés lors des étapes précédentes.

Malgré les nombreuses avancées dans le domaine de la synthèse de la parole en français depuis ces dernières années et les efforts importants pour la démocratisation de ce genre d'outil, force est de constater que l'utilisation actuelle de cette technologie est pour l'instant rudimentaire (enregistrement d'échantillons de parole supérieurs au mot), ce qui limite indubitablement ses champs d'applications. Les lacunes rencontrées sont souvent les erreurs de prononciation (dues au manque de robustesse des systèmes par rapport à la diversité des textes et aux ambiguïtés inhérentes au langage naturel) et le peu de naturel de la voix produite.

Les systèmes courants qui combinent une analyse syntaxique de surface et des règles morpho-phonétiques et syntactico-rythmiques ont des résultats qui restent insuffisants en partie à cause d'une analyse linguistique incomplète.

L'approche proposée utilise FIPS, notre analyseur basé sur un modèle Chomskyen de la grammaire générative qui fournit une analyse syntaxique détaillée aux modules de phonétisation et de génération de la prosodie que nous décrirons ici (voir figure 1). Ces modules fonctionnent sous l'hypothèse d'une analyse correcte et complète. En cas d'échec, FIPS fournit une analyse partielle constituée de la meilleure suite de syntagmes reconnus en se basant sur des heuristiques linguistiques. Durant la dernière étape, le signal acoustique est créé par le synthétiseur MBROLA de la Faculté Polytechnique de Mons [Dut93]. Nous décrivons ici les justifications linguistiques de la phonétisation et de la génération de la prosodie ainsi que la mise en œuvre technique.

2. PHONÉTISATION

2.1 Présentation

Une des étapes indispensable à une phonétisation correcte en français est la segmentation et l'analyse des unités lexicales. En effet il est important de pouvoir identifier la catégorie syntaxique d'un mot pour savoir comment le prononcer. Les homographes hétérophones sont relativement fréquents (comme *tous, plus* et *est*) et source d'inintelligibilité en cas de mauvais traitement comme l'illustre la phrase suivante:

(1)a. *Le soleil est à l'est.*

b. *est* se prononce[ɛst] en tant que substantif.

c. *est* se prononce [ɛ] en tant que verbe.

La plupart des systèmes de synthèse utilisent généralement des règles indépendantes du contexte ou sont basés sur des étiqueteurs types stochastiques pour identifier les parties du discours. Sans critiquer ces approches, notre but est de montrer qu'une analyse syntaxique détaillée globale permet tout naturellement de lever ce genre d'ambiguïté de prononciation. Ainsi dans l'exemple (2), une analyse locale se limitant à l'examen des deux mots suivants ne permet pas de déterminer la valeur verbale du mot *content.*

(2) *Les frères de l'enfant content et négligent la fin des histoires....*

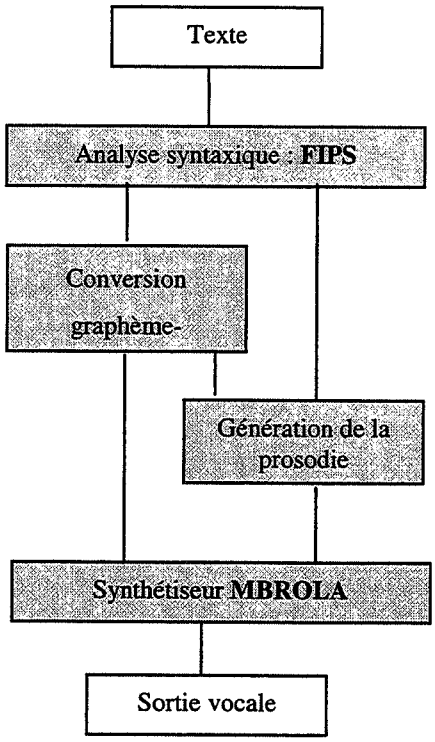


Figure1: Architecture du système de synthèse FipsVox

2.2 Phonétisation lexicale

Etant donné que l'analyseur syntaxique utilise un **lexique morpho-syntaxique**, l'approche d'une phonétisation par lexique est plus adapté qu'un système complet de règles. Notre lexique comporte 185 000 entrées et inclut une information phonétique, fréquentielle, ainsi que des traits sémantiques.

Un système expert de conversion graphèmes-phonèmes à base de **règles** ordonnées(environ 500) propose une solution alternative en cas de recherche infructueuse dans le lexique. Il est aussi utilisé lors d'ajouts d'entrées lexicales, pour lesquelles il intervient de façon interactive en proposant une forme phonétique pour ces nouveaux mots.

Les règles respectent le formalisme suivant (voir aussi [Div97]):

(3) GR:[PH]/CG+CD\CS;

où GR est la suite de graphèmes considérée, CG et CD les contextes optionnels à gauche et à droite, CS la condition syntaxique supplémentaire et finalement PH la transcription sur un ou plusieurs phonèmes de GR. Les CG et CD peuvent être de types : graphème unique, suite de graphèmes, début ou fin de mot, classes de graphèmes (qui permettent d'éviter la multiplication des règles) ou encore vide.

Dans les règles (4a) et (4b) qui suivent, la catégorie syntaxique permet de lever l'ambiguïté de prononciation des finales de mot *'ent'* (c.f. exemple 2) :

(4)a. 'ent' :[ə]/+'FinDeMot'\Verbe ;

b. 'ent' :[ɑ̃]/+'FinDeMot'\Adjectif ;

La sélection de la règle se fait d'abord en faveur de la séquence graphèmique la plus longue, puis selon l'ordre de déclaration des règles.

Dans le cas d'un mot inconnu, un devineur morpho-syntaxique propose certaines catégories. L'analyseur filtre ensuite ces propositions pour choisir la plus cohérente (avec le contexte syntaxique de la phrase), dont le phonétiseur prendra en compte (avec CS).

2.3 Phonétisation en contexte

Dans [Gau97] est montré l'intérêt intrinsèque de l'analyse syntaxique pour le traitement de la liaison du français tel que l'avait présenté Selkirk [Sel74] dans le cadre de la grammaire générative. Dans notre traitement des liaisons obligatoires et impossibles, la condition nécessaire de c-commande dans l'arborescence syntaxique entre le lieur et lié est associé à des conditions strictes de localité, d'appartenances catégorielle et morpho-phonologique (nombre de syllabes). Dans l'exemple (5) une analyse fine permet de lever l'ambiguïté d'appartenance catégoriel s'exerçant sur les mots *grand* et *arbitre*, qui en (5a) ne génère pas de liaison lorsqu'un substantif est suivi d'un verbe. Alors qu'en (5b) un substantif précédé d'un adjectif produit la liaison.

(5) a. Le grand‿arbitre la rencontre qui a lieu demain.

b. Ce grand arbitre la rencontre dans les vestiaires.

D'autre part, il faut noter l'existence du phénomène de dénasalisation qu'il faut prendre en compte pour une application de lecture naturelle. Ce phénomène concerne la transformation d'une voyelle non nasale suivie de la consonne nasale et est un sous-ensemble du phénomène de la liaison. C'est le cas de l'adjectif certain [sɛrtɛ̃] qui est prononcé [sɛrtɛn] dans *un certain enfant.*

Pour le traitement du e-muet (schwa) la simple utilisation de la fameuse règle des 3 consonnes, bien connue des enseignants de français langue seconde, permet de répondre à un traitement informatique de ce phénomène qui est extrêmement courant en français. Il faut, de plus garantir, une certaine flexibilité quant aux élisions de *e*, pour répondre aux variabilités de prononciation (origine linguistique, niveau de langue).

Indépendamment d'une analyse globale de la phrase, il est indispensable de pouvoir traiter certaines micro-grammaires de l'écrit tel que le format des dates, des numéro de téléphones, ou autres... Pour ce genre de tâche, l'utilisation de formalisme des automates à états finis semblent adéquates. Pour le moment FIPSVox permet de traiter la plupart des nombres écrits. Mais la composante syntaxique du système lui permet de faire beaucoup plus, puisqu'il peut phonétiser correctement les cas de

prononciations variables des nombres en fonction de leur contexte:

(6) a. *les 41 filles sont couchées.*

b. [lekarãteynfjsɔ̃kuʃe]

où le chiffre 1 doit se prononcer [yn] et pas [œ̃] car il s'accorde en genre avec le substantif *filles*.

3. GÉNÉRATION DE LA PROSODIE

Le module prosodique est basé sur une approche perceptuelle [Bea94] et génère le profil mélodique au moyen de mouvement dits standards. Le calcul de la durée segmentale combine une approche également perceptuelle et le modèle de Campbell basé sur le z-score [Cam93]. Quelques règles phonotactiques ont été ajoutées pour une meilleure répartition des accents, éviter des groupes prosodiques de longueur inadéquate (moins de 3 ou plus de 7) et autoriser une remise à zéro de la déclinaison (*baseline resetting*).

L'information lexicale et syntaxique, fournie par l'analyseur, est utilisée pour déterminer la dépendance des syntagmes de la phrase. Dans le cas classique des syntagmes nominaux successifs post-verbaux, il est indispensable de déterminer l'attachement du second pour générer une prosodie appropriée (un accent plus marqué sur la dernière syllabe de *gâteau* en (7b)).

(7) a. *Jean achète un gâteau de la crémière*

[TP[DP Jean][T' achète [VP [DP un [NP gâteau [PP de [DP la [NP crémière]]]]]]]]

b. *Jean achète un gâteau à la crémière.*

[TP[DP Jean][T' achète [VP [DP un [NP gâteau]] [PP à [DP la [NP crémière]]]]]]

De plus, l'analyse syntaxique permet de détecter et d'identifier des constructions particulières qui doivent être marquées prosodiquement, comme les énumérations, les incises et tout syntagme hors de sa position canonique : les syntagmes extraposés (8a), les dislocations à gauche (8b) et à droite (8c), les structures clivées (8d) (comme réponse à la question (8e) et non à la question (8f) pour laquelle la structure syntaxique est une relative simple) (cf. [Gol97] pour une description plus détaillée).

(8) a. *Jean aime bien les gâteaux, pas les biscuits.*

b. *Jean, il aime bien les gâteaux.*

c. *Jean les aime bien, les gâteaux*

d. *C'est les gâteaux que Jean aime bien.*

e. *Qu'est-ce que Jean aime, les gâteaux ou les biscuits ?*

f. *Qu'est-ce qu'il y a dans le four ?*

La génération de la prosodie de la phrase à synthétiser se réalise en deux phases : 1. durant le parcours de la structure syntaxique, les informations nécéssaires pour le calcul prosodique sont extraites, 2. à partir de ces informations minimales, plusieurs étapes calculent successivement les frontières des groupes prosodiques, la position et l'importance des accents puis le profil mélodique global et la durée de chaque phonème.

3.1 Parcours de la structure syntaxique

La représentation sous forme de structure arborescente de l'analyse syntaxique complète nous permet de connaître pour chaque mot non seulement sa catégorie grammaticale mais aussi un certain nombre de propriétés concernant cette catégorie. Pour le cas des adverbes par exemple, parmi les quatre types possibles (*sentential, interrogative, standard, négation*), ne sont considérés comme mots lexicaux (donc accentogènes) uniquement les adverbes des trois premiers types. Les adverbes de type *négation* sont classés comme mots outils (donc non accentogènes). Néanmoins, pour la catégorie *standard*, les adverbes qui ont pour valeur *intensity* (comme *très*) sont également marqués comme mot outil (ils prendront un accent dans des cas particuliers de focalisation relevant d'une analyse plus discursive comme dans *Est-ce que Jean est aimable ? Il est TRÈS aimable).*

Puis, une série de règles permet de positionner pour chaque mot, des étiquettes syntaxiques qui décrivent l'enchaînement et la dépendance des syntagmes successifs grâce à la structure syntaxique complète. Ces étiquettes déterminent la position et l'étendue des groupes nominaux (ainsi que la dépendance ou non de groupes nominaux successifs) (notés n ou N), des propositions subordonnées (P), des incises (I), des conjonctions et énumérations (E), de chaque groupe verbal (V) ainsi que de son sujet s'il n'est pas cliticisé (S), et finalement des syntagmes extraposés (X). La ponctuation ainsi que l'importance de la frontière syntaxique suivant le mot sont également pris en compte. L'exemple suivant montre les informations extraites de l'analyse syntaxique pour le calcul prosodique (étiquette, ponctuation, importance de la frontière syntaxique):

(9) a. *Tout-à-coup le chat de Jean dévore, goulûment, la souris de Sophie à pleines dents !*

[TP[AdvP tout-à-coup][DP le [NP chat [PP de [DP Jean]]]] [T' dévore [VP *{[AdvP[Adv' goulûment]]} [DP la [NP souris [PP de [DP Sophie]]]] [AdvP [PP à [NP[AP pleines][N dents]]]]]]]

Tout-à-coup(1) ,(S)(N) le chat (/N)(n) de Jean(4) (/n) (/S) (V) dévore (/V), (I) goulûment(2) (/I), (N) la souris (/N) (n) de Sophie (4) (/n) (N) à pleines dents (7) (/N) !

3.2 Génération de la prosodie

De cette représentation syntaxique simplifié et de la chaîne phonétique issue de la conversion graphème–phonème, se succèdent une série de sous-modules en cascade gérés par des règles appropriées :

- la syllabe ayant été choisie comme unité rythmique, une procédure de syllabification simple en trois règles découpe chaque mot phonétique en syllabes.
- les frontières de groupes sont calculés en fonction de la (ou des) étiquette(s) syntaxique(s) associées à chacun des mots ainsi qu'en fonction du nombre de syllabes des groupes prosodiques ainsi formés.
- les accents de groupes sont disposés selon une série de règles basées sur la position et la nature des frontières de groupes et de contextes particuliers comme les incises ou les syntagmes extraposés pour lesquels un traitement particulier est appliqué.
- la durée phonémique est déduite des accents. La méthode adoptée détermine dans un premier temps la durée syllabique grâce à une combinaison de travaux précédents qui prennent en compte : le nombre de phonèmes par syllabe, la position de la syllabe par rapport à l'accent suivant et l'importance de celui-ci (primaire,scondaire). Puis la durée de chaque phonème est estimée par le biais du *z-score* (allongement intrinsèque d'un ou plusieurs segments).
- finallement, le gabarit global de l'intonation est déterminé sur la base de la durée totale de l'élocution et d'un certain nombre de paramètres fixant les valeurs moyennes de la fréquence fondementale, du registre et de la déclinaison désirée.
- le profil mélodique final est généré à partir de la base de ce gabarit ainsi que de la position et de la nature des accents qui sont réalisés sous forme de mouvements standardisés.

4. Conclusion et Aspects Applicatifs

Dans cette contribution, nous avons présenté le système complet de synthèse de parole FIPSVox ainsi que les avantages de l'utilisation d'une analyse syntaxique détaillée dans le traitement des problèmes soulevés par la phonétisation et durant la génération de la prosodie. Une plus grande couverture de la variété textuelle et la phonétisation des noms propres sont les prochaines étapes envisagées. De plus une évaluation permettra de préciser de manière quantitative l'importance de l'apport de la syntaxe pour toutes les étapes de la synthèse.

En tant qu'application linguistique, nous avons tout naturellement orienté notre système de lecture de texte vers d'autres utilisations dans les domaines des technologies du langage. En effet, l'aspect résolument généraliste de notre système nous offre intrinsèquement de nombreuses perspectives. Plus concrètement l'intégration de notre application en sortie de systèmes de traduction ou en tant qu'outil d'apprentissage comme dans Safran (Système pour l'Apprentissage du Français), nous a fait réaliser l'importance de la mise au point d'un meilleur contrôle du synthétiseur aussi bien au niveau local que global. Par exemple dans les systèmes d'apprentissage il est important de pouvoir prolonger certaines syllabes pour insister sur un exercice de prononciation, mais d'un point de vue global il est aussi important d'avoir le choix entre plusieurs modes d'élocution (type familier ou soutenu par exemple). Dans FIPSVox, le choix du mode d'élocution se traduit pour le moment par un ajustement des règles d'élision du schwa, mais prochainement les liaisons et la prosodie devrait être aussi prise en compte.

Une version dynamique du synthétiseur est également disponible sur le serveur du LATL proposant la génération d'une phrase et l'illustration de la problématique par un jeu de phrases choisies. Les correspondance bilingues Anglais/Français de notre DicoVox, disponibles sur le serveur sont aussi prononcées grâce à notre système.

Bibliographie

[Bea94] Beaugendre F. (1994) "Une étude perceptive de l'intonation du français", PhD thesis, Université d'Orsay.

[Cam92] Campbell W.N. (1992) "Multi-level Timing in Speech" Ph.D. thesis. University of Sussex.

[Div97] Divay M. et Vitale A. (1997) "Algorithms for grapheme-phoneme translation for english and french: applications for database searches and speech synthesis", Computational Linguistics, Vol. 23, pp. 495-523.

[Dut93] Dutoit T. "The MBROLA Project", Info. on http://tcts.fpms.ac.be/synthesis/mbrola.html

[Gau97] Gaudinat A. et Wehrli E. (1997) "Analyse syntaxique et synthèse de la parole : Le projet FipsVox", TAL , Vol. 38, pp. 121-130.

[Gol97] Goldman J.Ph. et Wehrli E. (1997), « Deriving prosodic patterns from syntactic structures : the cases of extraposition, clefts and parentheticals », Intonation, ESCA Workshop.

[Lae91] Laenzlinger C. and Wehrli E. (1991) "FIPS : Un analyseur interactif pour le français", TA Informations, Vol. 32:2, pp. 35-49.

[Pas90] Pasdeloup V. (1990) "Modèles de règles rythmiques du français appliqué à la synthèse de la parole", PhD thesis, Université de Provence, 1990.

[Sel74] Selkirk E. (1984) "French liaison and the X convention", Linguistic inquiry, Vol. 5, pp. 573-590.

Une méthode de synthèse par copie pour piloter le synthétiseur de Klatt

Yves Laprie et Anne Bonneau
LORIA
BP 239
54506 Vandœuvre-lès-Nancy, FRANCE
laprie,bonneau@loria.fr

RÉSUMÉ

This paper presents a copy synthesis method to controling the Klatt synthetizer. Our method allows speech stimuli to be constructed very easily. We accepted the parallel branch of the Klatt synthetizer. After formants have been tracked, the amplitudes of the resonators are measured on a spectrum obtained by an algorithm derived from cepstral smoothing called "true envelope". This algorithm has the advantage of approximating harmonics very accurately. The analysis strategy of a speech signal is straightforward: the fundamental frequency is calculated so that voiced regions are known and the friction energy is set to the value of the spectral energy above 4000 Hz. Stimuli which have been created by means of this method have a timbre close to that of natural speech. This copy synthesis method is incorporated in our software for speech research called "Snorri". Therefore, the user has at his disposal a versatile tool for creating stimuli in the context of the Klatt synthetizer.

1. INTRODUCTION

Pour de nombreuses études de perception et de phonétique il est intéressant de disposer d'un système de synthèse avec lequel on puisse facilement contrôler les indices acoustiques. Un certain nombre de synthétiseurs à formants ont été proposés mais c'est incontestablement celui de Klatt [7] qui est le plus souvent utilisé car il est facilement disponible et s'est révélé capable de produire de la parole très proche de la parole naturelle.

Malgré son intérêt, le synthétiseur de Klatt est difficile à utiliser seul car il faut spécifier 39 paramètres (pour la version de 1980) pour chaque trame. Dans le cas d'un synthétiseur à partir du texte ces paramètres sont obtenus grâce à un ensemble de règles complexes et après qu'un jeu de paramètres a été déterminé pour les sons de base. Comme l'indique Klatt cette détermination est souvent le résultat d'essais-erreurs ce qui est très fastidieux. Ce travail préalable devient un obstacle plus important encore pour le phonéticien qui souhaite disposer de sons proches des sons naturels sans devoir consacrer trop de temps au réglage des paramètres de synthèse.

Pour que les signaux synthétisés soient jugés suffisamment naturels il semble donc séduisant d'extraire les paramètres à partir de la parole naturelle. Cette idée a donné lieu à des tentatives de « synthèse par copie », notamment celle de W. Holmes [6] qui repose sur un synthétiseur plus simple que celui de Klatt. Mais il n'existe pas, à notre connaissance, de système de synthèse par copie pour le synthétiseur de Klatt. Par rapport à des logiciels d'édition graphique des paramètres du synthétiseur de Klatt notre système a l'intérêt d'être intégré à un logiciel d'étude de la parole très complet (notre logiciel Snorri [9]) et de mettre à la disposition de l'utilisateur des outils de synthèse par copie. Dans cet article nous présentons ces outils.

2. CHOIX DU TYPE DE SYNTHÉTISEUR

Le synthétiseur de Klatt comporte deux parties : **(i)** l'une est une cascade de résonateurs dont il faut donner la fréquence et la largeur de bande, **(ii)** l'autre est un assemblage de résonateurs en parallèle, chaque résonateur étant défini par sa fréquence, sa largeur de bande et son amplitude.

Théoriquement, la branche en cascade est suffisante pour produire toutes les voyelles orales mais il faut la compléter par la branche parallèle pour produire les voyelles nasales et les consonnes. Cela signifierait donc que la synthèse par copie doive porter sur les deux branches. En fait, des expériences préliminaires, confirmant des travaux précédents effectués avec d'autres synthétiseurs à formants, nous ont montré qu'il est préférable de se limiter à la branche parallèle. Il est en effet possible de produire avec cette branche des voyelles ayant le même spectre que celles que produiraient la branche en cascade et seule la branche parallèle permet de traiter le cas des consonnes.

La branche parallèle possède un second avantage. Les formants de la branche en cascade sont définis par leur fréquence et leur largeur de bande. Mais la largeur de bande d'un formant est difficile à déterminer correctement. Il faut la mesurer pendant la phase fermée de la période fondamentale, c'est-à-dire quand le conduit vocal n'est pas couplé aux cavités sous-glottiques. Cela nécessite un marquage précis des instants de fermeture de la glotte ce qui n'est pas simple à mettre en œuvre (voir par exemple [2]) et peut donc conduire à un certain nombre d'erreurs. Or la largeur de bande contrôle aussi l'amplitude du formant et les erreurs sont donc doublement sensibles. Pour la branche parallèle, en revanche, l'amplitude et la largeur de bande sont deux paramètres indépendants. Certains synthétiseurs où les résonateurs sont assemblés en parallèle n'utilisent d'ailleurs que l'amplitude pourvu que la largeur de bande

soit fixée à une valeur acceptable pour chacun des formants.

C'est pourquoi nous avons choisi de ne travailler qu'avec la branche parallèle du synthétiseur de Klatt et que dans un premier au moins nous avons fixé la largeur de bande aux valeurs proposées par Mrayati et Guérin [11].

3. SYNTHÈSE PAR COPIE

La synthèse par copie consiste à déterminer à chaque instant la fréquence et l'amplitude des formants qui permettent de reproduire le plus fidèlement possible la parole originale. Les paramètres de base sont les fréquences des formants. Dans le logiciel que nous avons développé l'utilisateur a le choix entre l'extraction automatique des formants grâce à notre algorithme de suivi automatique [8] ou l'édition manuelle des trajectoires formantiques à l'aide du spectrogramme (l'utilisateur trace les formants directement sur le spectrogramme affiché à l'écran). Dans le second cas l'utilisateur peut d'ailleurs s'aider de deux outils : **(i)** un algorithme de recalage des trajectoires données par l'utilisateur sur les lignes de pics du spectrogramme calculé par prédiction linéaire ou lissage cepstral, **(ii)** un algorithme de lissage par B-spline avec un point de contrôle toutes les 80 ms qui est la durée moyenne d'une voyelle en français.

Il reste donc à déterminer les paramètres de la source et les amplitudes. Pour la source nous ne calculons que la fréquence fondamentale grâce à un algorithme dérivé de celui proposé par Martin [10].

La fonction de transfert de la parole produite par le synthétiseur est de la forme : $P(z) = S(z).H(z)$ où S représente la fonction de transfert de la source et H la contribution du conduit vocal et de la radiation aux lèvres : $H(z) = -H_1(z) + H_2(z) - H_3(z) + H_4(z) - H_5(z) + H_6(z)$ où $H_i(z)$ représente la fonction de transfert du résonateur associé au formant Fi.

Les paramètres d'amplitude des fonctions $H_i(z)$ sont obtenus par mesure à partir du spectre original et du spectre de la source. Il serait éventuellement possible de calculer $S(z)$ à partir de la forme analytique du signal temporel d'excitation $s(n)$. En fait, Klatt a ajouté un certain nombre de paramètres pour produire un signal d'excitation proche de la réalité ce qui donne une forme analytique assez lourde à $s(n)$. Comme il est possible d'utiliser un signal d'excitation extrait d'un signal naturel nous avons préféré calculer $S(z)$ directement à partir du fichier du signal $s(n)$ obtenu par synthèse ou un autre moyen.

L'influence d'un formant est en grande partie réduite au voisinage immédiat du formant et il est possible, au moins dans un premier temps, d'obtenir le paramètre d'amplitude par simple mesure. Étant donné la fonction de transfert du formant Fi : $H_i(z) = \frac{A_i(1-B_i-C_i)}{1-B_iz^{-1}-C_iz^{-2}}$ où B_i et C_i sont des constantes réelles définies à partir de la fréquence et de la largeur de bande du formant Fi, A_i est donnée par $A_i = 20log_{10}|P(z_i)| - 20log_{10}|S(z_i)| - 20log_{10}|H(z_i)|$ où z_i est la fréquence complexe correspondant à la fréquence du formant F_i. Le signal de source, sauvegardé après synthèse, est pré-accentué pour affaiblir l'influence des formants $F2$ à $F6$ en basse fréquence. En revanche le signal d'excitation de $F1$ ne doit pas être pré-accentué et H_1 est donc de la forme $H_1(z) = \frac{A_1(1-B_1-C_1)}{(1-B_1z^{-1}-C_1z^{-2})(1-z^{-1})}$

4. RÉGLAGE DES AMPLITUDES

Le réglage des amplitudes est simple à condition d'évaluer correctement l'amplitude du spectre de la parole naturelle et celui de l'excitation. C'est en fait l'une des difficultés majeures de la synthèse par copie. Les formants sont souvent extraits à partir d'un spectre calculé par prédiction linéaire. Mais les faiblesses de ce type d'analyse spectrales sont connues : les formants sont « attirés » par les harmoniques intenses et les sons autres que les voyelles sont mal modélisés. À cela s'ajoute une forte imprécision sur la mesure des amplitudes qui est souvent de l'ordre de 10 dB et le fait que les erreurs commises sur le spectre de parole originale et sur le spectre du signal d'excitation peuvent s'ajouter.

Le lissage cepstral présente moins de faiblesses mais il est assez sensible à la position de la fenêtre d'analyse et le spectre lissé peut être nettement au-dessous des harmoniques. Nous avons donc retenu une méthode dérivée du lissage cepstral et introduite par Halle [3] appelée « enveloppe vraie ». Son principe est de partir du lissage cepstral classique et de le corriger itérativement en réduisant la contribution des valeurs spectrales situées au-dessous des pics spectraux (notamment les harmoniques). Cette méthode, bien qu'un peu lourde en calculs puisqu'il faut calculer deux transformées de Fourier par itération donne d'excellents résultats tant sur le spectre de la parole originale que sur le spectre d'excitation (voir Fig. 1).

4.1. Amplitudes des premiers harmoniques

À ce stade du réglage des paramètres la parole synthétisée est tout à fait acceptable. La différence essentielle porte sur les premiers harmoniques qui sont en général plus forts dans la parole naturelle. La parole synthétisée a donc un timbre moins grave mais J.N. Holmes [4] assure cependant que la contribution de ces harmoniques sur la perception des sons de la parole est faible. Pour obtenir de la parole de bonne qualité il propose de contrôler les premiers harmoniques par l'intermédiaire d'un formant de largeur de bande assez forte (150 Hz) et d'une fréquence de 200 Hz. Ce formant est aussi utilisé pour la synthèse des voyelles nasales. Nous avons donc utilisé le formant FN du synthétiseur de Klatt pour modifier l'amplitude des premiers harmoniques. Son amplitude est réglée de la même manière que celle de $F1$ car le signal d'excitation ne doit pas être préaccentué.

4.2. Stratégies d'analyse de la parole

La stratégie d'analyse que nous avons adoptée pour nos premiers essais a donné de bons résultats malgré sa simplicité. Une fois la fréquence fondamentale calculée, l'intensité du voisement est fixée à 60 dB pour toutes les régions voisées et celle de la friction est fixée à 60 dB pour tous les segments non voisés (fricatives sourdes et occlusives sourdes). Le réglage automatique des amplitudes assure

que les niveaux d'énergie des formants sont corrects.

Cas des fricatives Dans les segments fricatifs les premiers formants sont donc considérablement affaiblis et les formants $F4$ à $F6$ correspondent au bruit de friction. Pour déterminer les formants $F4$ à $F6$ nous avons utilisé une fenêtre temporelle très longue (48 ms) pour les analyses spectrales qui évite d'obtenir des trajectoires très instables en fréquence.

La stratégie d'analyse que nous venons de décrire doit être cependant légèrement corrigée pour tenir compte des fricatives voisées pour lesquelles il doit y avoir simultanément voisement et friction. Nous fixons donc l'amplitude de la friction à la valeur de l'énergie de la parole au-dessus de 4000 Hz, et cette composante est toujours prise en compte pour le calcul de l'excitation. La modélisation des fricatives voisées est cependant légèrement insuffisante pour deux raisons. D'abord il faudrait pouvoir contrôler plus finement la proportion de friction et de voisement dans l'excitation comme l'a fait J.N. Holmes dans le synthétiseur JSRU [5] où il est possible de régler le niveau de la friction pour chacun des formants. Ensuite il faudrait disposer d'une méthode d'analyse permettant de localiser le bruit dans la parole.

Cas des occlusives Jusqu'à présent nous n'avons pas développé d'analyse spécifique pour les occlusives car nous avons jugé les résultats actuels satisfaisants. Le seul point sur lequel nous reviendrons est le début de la barre d'explosion des occlusives qui nécessite l'utilisation du paramètre « by-pass ». Ce paramètre laisse passer toute l'énergie d'excitation et permet d'obtenir un spectre diffus nécessaire pour les occlusives labiales et dentales. Pour localiser la barre d'explosion nous utiliserons la méthode de segmentation des bursts que nous avons proposée dans [1] pour l'analyse et l'identification des occlusives.

4.3. Optimisation de la largeur de bande

Jusqu'à présent nous avons fixé la largeur de bande à une valeur par défaut qui dépend du formant. Cela nous permet de régler les amplitudes par simple mesure mais n'assure pas un approximation très précise du spectre original en particulier quand deux formants sont proches. Nous avons donc ajouté une procédure d'optimisation qui part des paramètres mesurés sur le spectre et qui minimise l'écart entre le spectre d'origine et le spectre de synthèse. L'optimisation peut porter sur tous les paramètres (fréquence, amplitude et largeur de bande du résonateur) mais il n'est cependant pas possible de les optimiser tous en une seule fois car le début du spectre n'est pas toujours correctement modélisé par le synthétiseur de Klatt ce qui fait parfois échouer l'optimisation (qu'elle utilise le gradient comme la méthode de Davidson-Fletcher-Powell ou non comme la méthode de Powell). Il faut donc dans certains cas fixer les valeurs de certains paramètres (souvent les fréquences de F1 et FN) et optimiser le spectre sur plusieurs régions indépendantes.

5. EXPÉRIMENTATIONS

Nous construisons actuellement une base de sons de synthèse comprenant des logatomes ainsi que quelques phrases. Cette base sera utilisée pour des études de perception. Les sons synthétisés sont créés à partir de la parole naturelle et, plus précisèment, à partir des enregistrements d'un même locuteur.

Avant d'expliquer comment les stimuli ont été créés, nous présentons brièvement comment fonctionne le logiciel de pilotage. Ce pilotage offre à l'expérimentateur le choix entre le calcul automatique des paramètres (fait par Snorri) et leur tracé manuel effectué directement sur le spectrogramme. Les deux fonctions sont du reste complémentaires puisqu'il est possible de corriger manuellement les résultats des calculs automatiques ou de recaler par calcul automatique les tracés manuels. Chaque paramètre peut également avoir une valeur par défaut. Le logiciel de pilotage est implanté dans le logiciel d'édition du signal Snorri qui offre de nombreuses possibilités de calcul et d'affichage des principaux indices de la parole.

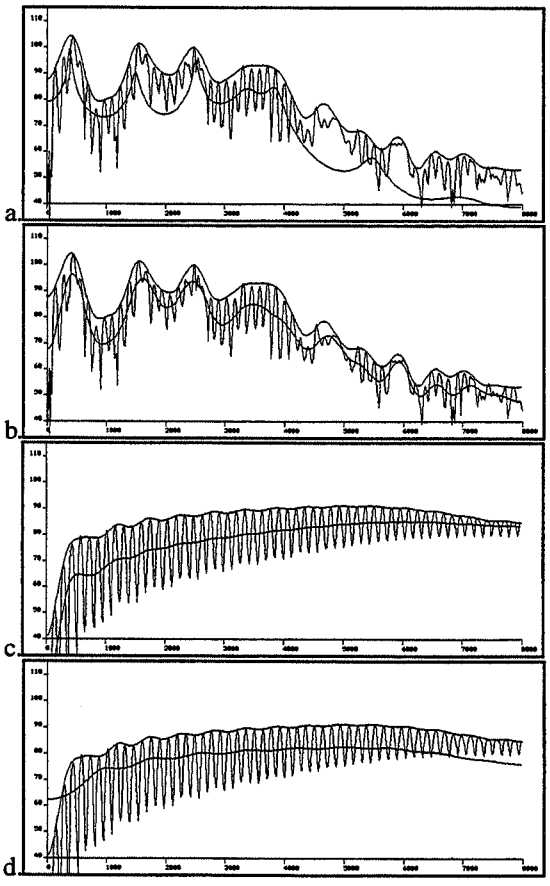


FIG. 1 – *Comparaison de « l'enveloppe vraie » à la prédiction lineaire (a et c) et au lissage cepstral (b et d) pour un spectre de parole (a et b) et un spectre du signal d'excitation (c et d). Sur chacune des figures apparaissent le spectre obtenu par transformée de Fourier bande étroite, le spectre obtenu par « enveloppe vraie » et l'analyse spectrale à comparer.*

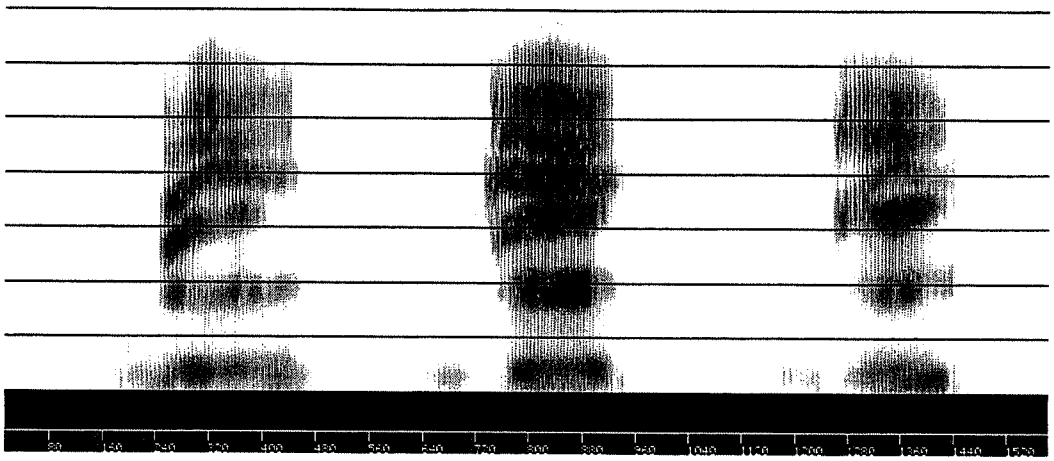


FIG. 2 – *Spectrogramme des signaux /bi/ /di/ /gi/ produits par synthèse par copie*

Nous avons pour le moment créé quelques logatomes, très difficiles à différencier les uns des autres : /bi,di,gi/. Un premier jeu de stimuli a été créé avec comme seuls paramètres variables les fréquences et les amplitudes de formants, ainsi que la fréquence du fondamental. Nous avons donné à tous les autres paramètres une valeur par défaut identique pour tous les sons. Signalons néanmoins que le paramètre de friction a, par défaut, des valeurs différentes pour les voyelles et pour les consonnes. Les fréquences formantiques et les amplitudes ont été évaluées par la méthode de « l'enveloppe vraie ». La figure Fig. 2 montre le spectrogramme de ces logatomes synthétisés. Les syllabes synthétisées sont non seulement très intelligibles mais encore ont un timbre très proche de celui de la parole naturelle. Notre prochain travail consistera à essayer d'améliorer encore l'intelligibilité de ces stimuli, en particulier celle des occlusives, et à augmenter la taille de notre base afin qu'elle contienne les logatomes CV combinant tous les sons du français.

La copie par synthèse permet de donner aux sons synthétiques une qualité vocale et une intelligibilité proches de celles de la parole naturelle. En effet, les légères variations de la fréquence fondamentale et des fréquences des formants et, en général, toutes ces « imperfections » caractéristiques de la voix humaine, sont reproduites par cette technique. Elle reproduit également les liens complexes unissant les différents paramètres vocaux.

6. CONCLUSION ET PERSPECTIVES

L'intérêt du pilotage présenté ici repose sur sa grande flexibilité dans le calcul des paramètres, évoquée au chapitre précédent, et sur son couplage avec le logiciel d'édition du signal Snorri.

Nous nous sommes efforcés de ne pas modifier le synthétiseur de Klatt pour que les stimuli que nous obtenons soient utilisables en dehors de notre logiciel d'étude de la parole. Malgré tout, nous reviendrons sans doute sur ce choix notamment dans le cas des fricatives voisées pour lesquelles il faudrait pouvoir contrôler les paramètres de voisement et de friction du signal d'excitation en fonction des formants.

RÉFÉRENCES

[1] Bonneau (A.), Coste (S.), Djezzar (L.) et Laprie (Y.). – Two Level Acoustic Cues for Consistent Stop Identification. *In : Proceedings International Conference on Spoken Language Processing*, pp. 511–514. – Banff (Alberta, Canada), October 1992.

[2] Cheng (Y. M.) et O'Shaunessy (D.). – Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, n 12, December 1989, pp. 1805–1815.

[3] Halle (P.). – Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et la détection du pitch. *In : Actes du séminaire "Traitement du signal de parole"*, pp. 83–93. – Paris, 1983.

[4] Holmes (J. N.). – Formant synthetisers: cascade or parallel ? *Speech Communication*, vol. 2, 1983, pp. 251–273.

[5] Holmes (J. N.). – A parallel formant synthetizer for machine voice output. *In : Computer Speech Processing*, éd. par Fallside (F.) et Woods (W. A.). – Prentice Hall International, 1985.

[6] Holmes (W. J.). – Copy synthesis of female speech using the JSRU parallel formant synthetiser. *In : Proceedings of European Conference on Speech Technology*, pp. 513–516. – Paris, France, September, 1989.

[7] Klatt (D.H.). – Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Amer.*, vol. 67, n3, March 1980, pp. 971–995.

[8] Laprie (Y.) et Berger (M.-O.). – Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, vol. 19, n4, October 1996, pp. 255–270.

[9] Laprie (Y.) et Mercier (L.). – Un environnement logiciel pour un atelier phonétique. *In : Actes 20èmes Journées d'Etude sur la Parole*, pp. 209–214. – Trégastel, Juin 1994.

[10] Martin (Ph.). – Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne. *In : Actes des 12èmes Journées d'Etudes sur la Parole*, pp. 223–232. – 1981.

[11] Mrayati (M.) et Guérin (B.). – Étude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec perte. *Revue d'Acoustique*, no36, 1976.

Un générateur de prosodie "tout automatique"

Fabrice Malfrère[1,*], Thierry Dutoit[*] et Piet Mertens[*,**]

[*]Faculté Polytechnique de Mons, Service de Traitement du Signal
Boulevard Dolez 31 7000 MONS (Belgique)
[**]K.U.Leuven, Département de Linguistique
Blijde Imkomstraat 21 3000 Leuven
Tél.: ++32 65 37 41 33 - Fax: ++32 65 37 41 29
e-mail: malfrere@tcts.fpms.ac.be - http://tcts.fpms.ac.be/synthesis

ABSTRACT

Text-to-Prosody systems based on the use of prosodic databases extracted from natural speech will be a key point for further development of new Text-to-Speech systems. This paper describes a first attempt to generate the rhythm and the intonation of a French written text based on the use of prosodic databases. The system is based on a very crude chinks 'n chunks prosodic phrasing algorithm and on a prosodic analysis of a natural speech database. The system has been tested on single sentences and news paragraphs. The prosody seems to be acceptable most of the time.

1. INTRODUCTION

Les systèmes de synthèse de la parole à partir du texte sont de nos jours capables de produire une parole synthétique de grande qualité à la fois naturelle et intelligible [Dut97]. Pour obtenir de tels résultats, le rythme et l'intonation utilisés par le synthétiseur doivent être copiés sur ceux d'un locuteur humain prononçant le même texte. Des outils permettant de réaliser cette transplantation prosodique, tels que *PROTRAN* [Coi94] ou *MBROLIGN* [Mal97] permettent d'extraire automatiquement le rythme et l'intonation d'un signal de parole. Le système *MBROLIGN* est d'ailleurs disponible gratuitement pour utilisation non commerciale et non militaire, à l'adresse http://tcts.fpms.ac.be/synthesis/mbrolign. Néanmoins, ce type de synthèse par transplantation de prosodie ne permet pas d'envisager la synthèse d'un texte quelconque.

Dans le domaine de la synthèse à partir d'un texte, les générateurs automatiques de prosodie ne permettent pas, pour l'instant, d'obtenir une parole synthétique suffisamment naturelle. Dès lors, on a vu exploser depuis quelques années le nombre de recherches portant sur la génération automatique de la prosodie. Parmi les développements les plus récents de ce domaine, on trouve un intérêt grandissant pour les techniques d'apprentissage automatique telles que les réseaux de neurones [Tra92], les arbres de régression et de classification (comme dans [Hir94], les modèles de Markov [Boe93], les techniques de régression linéaire [Bla96] ou d'autres méthodes stochastiques [Mob93]).

L'ensemble de ces techniques se base sur l'analyse de corpus de parole préalablement étiquetés prosodiquement. La méthode présentée dans cette communication se trouve directement dans la lignée de ces dernières en ce qui concerne la génération des paramètres acoustiques (durées phonétiques et courbes intonatives). Néanmoins, ces paramètres ne sont pas directement dérivés du texte mais d'une description linguistique de l'intonation, à l'instar de la plupart des méthodes citées précédemment hormis [Bla96].

Dans la section suivante, les diverses étapes du système de génération de la prosodie sont décrites. La section 3 donne ensuite quelques commentaires sur les résultats obtenus et sur les problèmes posés par ce type d'approche. La section 4 tire des conclusions quant à l'utilisation de bases de données prosodiques pour la création de générateurs automatiques de prosodie à partir d'un texte. Quelques pistes permettant d'améliorer le système actuel sont également envisagées.

2. GÉNÉRATION DE LA PROSODIE

Le système de génération de la prosodie se décompose en 2 étapes:

- la génération d'une description symbolique de la prosodie à partir du texte et de sa syntaxe;
- le passage de cette représentation abstraite à une description acoustique de la prosodie comportant la génération des durées phonétiques et celle des contours mélodiques.

La phonétisation ainsi que la structure syntaxique du texte sont fournies au générateur prosodique par le système de phonétisation et d'analyse morpho-syntaxique LIPSS (*LI*nguistic *P*rocessing for *S*peech *S*ynthesis) [Dut93]. La sortie du générateur de prosodie est ensuite fournie au synthétiseur MBROLA ([Dut93], http://tcts.fpms.ac.be/synthesis) afin d'obtenir une sortie vocale et de permettre une évaluation du moins qualitative de la prosodie ainsi obtenue.

[1] Recherche subsidiée par le F.R.I.A. (**F**onds pour la **R**echerche dans l'**I**ndustrie et l'**A**griculture).

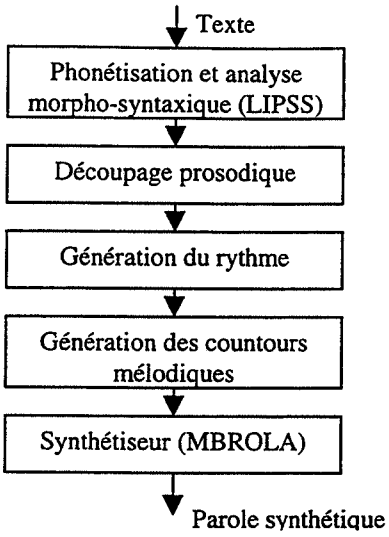


Figure 1 : Système de synthèse de la parole

2.1 Description symbolique de la prosodie

Dans cette première phase, le texte est décomposé en 'groupes' prosodiques selon un algorithme basé sur la détection des mots grammaticaux du texte (chinks ' n chunks [Lib92]).

Table 1: Caractère clitique ou non des mots

Nature	Clitique	Non clitique
Verbe		X
Infinitif		X
Nom		X
Adjectif		X
Adverbe		X
Déterminant	X	
Pronom	(X)	(X)
Préposition	X	
Conjonction de coordination	X	
Conjonction de subordination	X	
Participe présent		X
Participe passé		X

Les mots grammaticaux (aussi appelés outils ou clitiques) ne sont pas susceptibles de recevoir l'accent final. La position des accents finaux, et par là la délimitation des groupes intonatifs au sens de P. Mertens [Mer87], est alors déduite selon la règle suivante: si un mot clitique suit un mot non clitique, alors ce dernier reçoit l'accent final et marque la fin d'un groupe. Un accent final particulier est aussi attribué au dernier mot non clitique de la phrase ainsi qu'avant toute marque de ponctuation. Le tableau 1 reprend l'ensemble des catégories grammaticales générées par LIPSS et leur caractère clitique ou non. Des accents finaux supplémentaires sont ajoutés de manière à supprimer les suites de syllabes inaccentuées trop longues (ce qui explique l'accent final sur canard)

Un caractère accentué ou non est ensuite attribué à chaque syllabe en tenant compte du fait que seule la dernière syllabe du groupe prosodique reçoit un accent final. La dernière syllabe du mot terminant la phrase reçoit un accent particulier (L-L-, issu du modèle tonal de l'intonation de P. Mertens).

C	NC	NC	NC	C	NC
NA	NA	AF	AF	NA	AF
Le	petit	ca nard	a pprend	à	na ger .
NA	NANA	NA AF	NA AF	NA	NA L-L-
(		)	(		)

La figure 2 donne un exemple de formation des groupes prosodiques.

Figure 2 : Formation des groupes prosodiques
(C = clitique, AF = accentué, NA = non accentué, L-L- = fin de phrase)

2.2 De la prosodie symbolique à la prosodie acoustique

Une description symbolique relativement simple de l'intonation étant ainsi obtenue, sa conversion en durées phonétiques et en courbes mélodiques doit être réalisée. Comme annoncé dans l'introduction, cette phase de la génération de la prosodie se base sur l'analyse de corpus de parole étiquetés prosodiquement, c'est-à-dire pour lesquels on dispose de:

- la segmentation phonétique et syllabique;
- la fréquence fondamentale durant les zones voisées;
- le marquage prosodique symbolique.

Ces informations peuvent être obtenues par un étiquetage soit manuel soit automatique. Les deux cas ont été envisagés.

Un premier corpus monolocuteur d'une durée d'environ 10 minutes (2 articles du journal Le Soir) a été étiqueté phonétiquement et prosodiquement de manière automatique. L'étiquetage phonétique a été réalisé au moyen du système *MBROLIGN* [Mal97], la fréquence fondamentale extraite par l'algorithme MBE [Gri88] et ensuite lissée par l'intermédiaire d'un filtre médian d'ordre 5. En ce qui concerne le marquage prosodique, il a simplement été obtenu en appliquant l'algorithme décrit à la section 2.1 au texte composant le corpus (CORPUS1).

Un second corpus (extrait de BDTONS [Geo95]) monolocuteur étiqueté manuellement suivant la notation tonale de P. Mertens [Mer87] a également été utilisé (CORPUS2). Ce dernier est aussi d'une durée avoisinant les 10 minutes.

Génération du rythme

Les durées phonétiques, qui caractérisent le rythme, sont calculées par un ensemble de règles dérivées automatiquement d'un corpus de parole étiqueté.

Ces règles permettent d'affecter au phonème une durée en fonction du caractère accentué ou pas de la syllabe à laquelle il appartient ainsi que du nombre de phonèmes composant la syllabe. Un exemple de règles est donné dans la table 2 pour le phonème / *a* /.

Table 2 : Durée du phonème / *a* / en fonction de l'accent et de la taille de la syllabe (en phonèmes)

/ *a* /	1	2	3	> 3
AF (CORPUS1)	211 ms	152 ms	107 ms	60 ms
NA (CORPUS1)	96 ms	79 ms	73 ms	55 ms
AF (CORPUS2)	198 ms	136 ms	126 ms	110 ms
NA (CORPUS2)	72 ms	68 ms	62 ms	56 ms

En ce qui concerne les pauses, elles sont, pour l'instant, uniquement assignées en fonction de la ponctuation du texte et leur durée est fonction du signe de ponctuation.

Génération des contours mélodiques

La conversion de la description symbolique de l'intonation vers sa représentation numérique sous forme de marqueurs de F_0 se fait par comparaison des motifs prosodiques générés avec ceux se trouvant dans une base de données générée automatiquement à partir d'un corpus.

Les motifs prosodiques sont formés par l'ensemble des marqueurs prosodiques (NA,AF,L-L-) partant d'une syllabe accentuée jusqu'à la suivante incluse. Les motifs se recouvrent donc sur la syllabe accentuée. Une concaténation (au sens de la synthèse par diphones) des motifs se fait alors au niveau des syllabes accentuées.

Si on reprend l'exemple de la figure 2, il contient deux motifs prosodiques qui sont : NA NA NA NA AF et AF NA NA NA NA L-L- .

La base de données est constituée de l'ensemble des motifs apparaissant dans le corpus d'apprentissage ainsi que de la position du motif dans la phrase. A chaque entrée est associée une suite de valeurs de F_0 (5 par syllabe).

Les contours mélodiques sont déterminés en recherchant dans ce répertoire de motifs, le motif le plus proche de celui rencontré lors de la synthèse. Cette recherche se fait tout d'abord en essayant de trouver un motif dans la base qui soit exactement le même que celui à réaliser et qui se trouve dans la même position dans la phrase. Si ce motif n'existe pas, on laisse tomber la contrainte sur la position, puis sur le nombre de syllabes non accentuées composant le motif. L'entrée adéquate étant trouvée dans la base de données, il suffit alors d'appliquer les marqueurs de F_0 sur les parties voisées des syllabes, support segmental du motif. Les accents finaux, points de jonction des motifs, sont réalisés en prenant la moyenne des valeurs de F_0 terminant un motif et de celles commençant le motif suivant.

Enfin, lors de la recherche, afin d'assurer une plus grande variabilité, une contrainte supplémentaire est ajoutée: la réalisation de deux motifs se succédant ne peut correspondre à la même entrée dans le dictionnaire des motifs prosodiques.

Une fois l'ensemble du contour déterminé, on vérifie que celui-ci ne contient pas de variations de F_0 intra- et inter- syllabique de trop grande amplitude. Si c'est le cas elles sont écrêtées par un seuil de variation inter-syllabique et un seuil intra-syllabique.

L'ensemble du système de génération de la prosodie a été implémenté en utilisant une structure de données particulière appelée MLC (*M*ulti *L*ayer *C*ontainer, cf. Figure 3), permettant une organisation aisée des différentes couches de données d'un synthétiseur de la parole à partir du texte. Cette structure de données s'inspire des MLDS (*M*ulti-*L*evel *D*ata *S*tructure) de [Lee93].

3. RÉSULTATS

Les premiers résultats obtenus sont très encourageants tant pour le rythme que pour les contours mélodiques. Des tests ont été effectués sur des textes de longueur variable allant de la simple phrase à l'article de journal.

En ce qui concerne le rythme, on remarque que pour les phrases longues et sans marque de ponctuation, l'absence de pauses se ressent de manière importante comme on pouvait s'y attendre vu la méthode d'assignation des pauses employée.

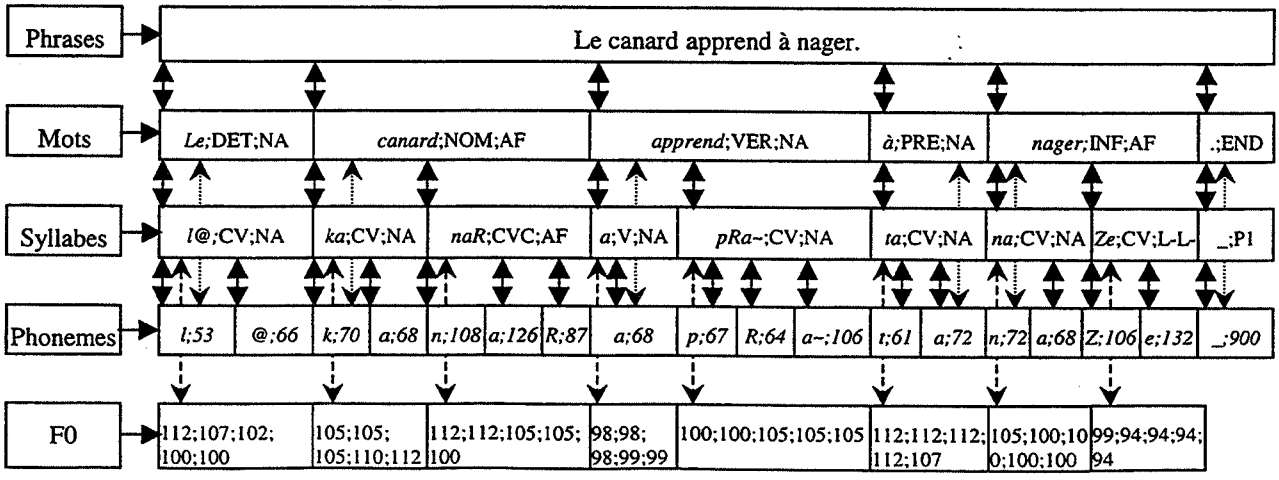


Figure 3: MLC (*M*ulti *L*ayer *C*ontainer) pour la phrase *"Le canard apprend à nager."*

Les quelques problèmes que l'on retrouve dans les contours mélodiques sont pour la majorité dus à une mauvaise accentuation. Ce problème provient essentiellement de la mauvaise assignation de certaines frontières prosodiques par le système, probablement à cause de la simplicité de l'approche.

Des tests d'écoute sont en cours. Quelques exemples de synthèse de la parole utilisant ce générateur prosodique peuvent être trouvés à l'adresse http://tcts.fpms.ac.be/synthesis/prosody.

4. CONCLUSIONS

Le générateur de prosodie à partir du texte décrit dans ce papier montre qu'il est possible d'utiliser des corpus de parole étiquetés automatiquement et d'obtenir une prosodie dont la qualité est très proche de celle qui est obtenue par les mêmes méthodes travaillant sur des corpus étiquetés manuellement.

Cette approche a le grand avantage de permettre de changer rapidement de locuteur de référence, et de fournir ainsi de multiples générateurs de prosodie ayant chacun un '*style*' particulier.

Pour améliorer la qualité de la prosodie obtenue à l'aide de ce système, plusieurs voies sont actuellement expérimentées :

- utiliser un modèle linguistique de représentation symbolique de l'intonation plus évolué tel que celui de P. Mertens [Mer97];
- ajouter une programmation dynamique au sein de la recherche du motif dans le dictionnaire afin de trouver la suite de motifs qui se concatènent le mieux à la manière de ce qui se fait en synthèse par concaténation d'unités non uniformes;
- affiner l'assignation des pauses, pour le moment uniquement basée sur la ponctuation;
- augmenter la taille du corpus de parole servant à créer les bases de données afin d'autoriser une plus grande variabilité de la prosodie générée.

5. REMERCIEMENTS

Les auteurs sont particulièrement reconnaissants envers Michel Bagein et à Alain Ruelle pour leur implémentation des MLCs et leurs nombreux conseils lors de leur mise en œuvre dans cette application.

BIBLIOGRAPHIE

[Bla96] Black A.W. & Hunt A.J. (1996), " Generating F_0 contours from ToBI labels using linear regression ", Proceedings of ICSLP'96, pp. 1385-1388.

[Boe93] Boëffard O. (1993), " Segmentation Automatique d'Unités Acoustiques pour la Synthèse de la Parole ", Thèse de Doctorat, Université de Rennes I.

[Coi94] Van Coile B., van Tichelen L., Vostermans A., Wang J.W. & Staessens M. (1994), " PROTRAN: a Prosody Transplantation Tool for Text-to-Speech Applications ", Proceedings of ICSLP'94.

[Dut93] Dutoit T. (1993), " High-Quality Text-to-Speech Synthesis for the French Language". Thèse de Doctorat. Faculté Politechnique de Mons.

[Dut97] Dutoit T. (1997), " An Introduction to Text-to-Speech Synthesis ", Kluwer Academic Publishers.

[Geo95] Geoffrois E. (1995), " Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole ", Thèse de doctorat, Université de Paris XI.

[Gri88] Griffin D. & Lim J. (1988), " Multiband Excitation Vocoder ", IEEE Transaction on ASSP, pp. 1123-1235, 1988

[Hir94] Hirschberg J. & Prieto P. (1994), " Training Intonational Phrasing Rules Automatically for English and Spanish Text-to-Speech ", Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pp. 159-162.

[Lee93] van Leeuwen H.C. & te Lindert E. (1993), " Speech Maker: a flexible and general framework for text-to-speech synthesis, and its application to Dutch ", Computer Speech and Language, pp. 149-167.

[Lib92] Liberman M.J. & Church K.W. (1992), " Text Analysis and Word Pronunciation in Text-to-Speech Synthesis ", in Advances in Speech Signal Processing, Furui S., Sondhi M.M., eds., Dekker, pp. 791-831.

[Mal97] Malfrère F. & Dutoit T. (1997), " High Quality Speech Synthesis for Phonetic Speech Segmentation ", Proceedings of the Eurospeech'97, pp. 2631-2634.

[Mob93] Möbius B., Pätzold M. & Hess W. (1993), " Analysis and synthesis of German F_0 contours by means of Fujisaki's model ", Speech Communication, pp. 53-61.

[Mer87] Mertens P. (1987), " L'intonation du français. De la description linguistique à la reconnaissance automatique ", Thèse de Doctorat, Katholieke Universiteit Leuven.

[Mer97] Mertens P. (1997), " De la Chaîne Linéaire à la Séquence de Tons ", Traitement automatique des langues, Vol. 38, n°1.

[Tra92] Traber C. (1992), " F0 Generation with a Database of Natural F0 Patterns and with a Neural Network ", in Talking Machines : Théories, Models, and Designs, G. Bailly and C. Benoît, eds. North Holland, pp. 287-304.

Utilisation d'une structure arborescente pour une hiérarchisation fine des règles de transcription graphème-phonème

Michel MOREL, Anne LACHERET-DUJOUR
Laboratoire de Linguistique ELSAP - UPRES-A 6047
Université de Caen, Esplanade de la Paix, 14032 CAEN CEDEX
e-mail : morel@elsap.unicaen.fr, lacheret@elsap.unicaen.fr

Abstract

The grapheme-phoneme transcription module which is presented here is one element of the text-to-speech Kali project which was launched in 1995 at Caen University and which groups together four partners : the research laboratories Elsap and Greyc, the firm Electrel, present since 1981 on the voice synthesis market, and the Club Micro Son, association which unites blind users of computers and provides research, development, training and equipment.

It is pointed out in this paper that a tree-diagram structure for the writing of rules has advantages from a practical viewpoint (readability, upgrading, success rate and transcription speed) as well as from that of the theoretical validity.

1. Introduction

Dans l'état actuel des recherches en transcription graphème-phonème, se pose le problème épineux de la gestion d'un jeu de règles de plus en plus volumineux et pourtant nécessaire pour une phonétisation de haute qualité : un transcripteur aujourd'hui doit pouvoir traiter en temps réel tout type de texte incluant bien sûr titres, dates, nombres, sigles, abréviations, noms propres et mots d'emprunt.

Après un bref état des lieux, nous souhaitons mettre en lumière ici les avantages qu'offre une structure de règles arborescente, tant d'un point de vue pratique que pour la cohérence théorique des sous-ensembles créés. Une évaluation du système est enfin proposée qui renseigne sur la vitesse d'exécution et sur le taux de transcription correcte.

2. Situation du problème

Les faibles possibilités des systèmes informatiques ont orienté la transcription graphème-phonème à ses débuts vers la conception de règles peu coûteuses en mémoire : Le Cornec [Lec72], Divay et Guyomard [Div77], Léty [Lét80], Prouts [Pro80]. Les résultats étaient déjà très acceptables car, dans la plupart des langues, quelques centaines de règles suffisent à décrire 99 % des prononciations existantes. La transcription graphème-phonème est typique du phénomène de décroissance exponentielle de l'efficacité des règles, bien connu en modélisation : peu de règles suffisent à décrire une grande majorité des cas, puis de plus en plus de règles sont nécessaires pour une amélioration de moins en moins sensible. Pour la phonétisation, il semble même qu'un doublement du nombre de règles ne suffise pas à diviser par deux le taux d'échec.

Il est donc devenu nécessaire d'améliorer la stratégie de transcription : d'un côté augmenter l'efficacité des règles, par exemple en créant davantage de catégories, de l'autre, utiliser des lexiques. Parallèlement, des efforts ont été faits pour faciliter la maintenance des règles en les rendant déclaratives (Divay [Div84]) ainsi que pour structurer les données sur des critères linguistiques (Catach [Cat84]). La plupart des systèmes actuels utilisent à la fois des règles et des lexiques, en proportions variables : Aubergé [Aub85], Bailly et Alissali [Bai92], Belrhali, Libert, Aubergé et Boë [Bel92], Belrhali [Bel95], Boula de Mareuil [Bou97].

La priorité est parfois donnée aux lexiques (Laporte [Lap88], Tihoni et Pérennou [Tih91]) et des travaux de recherche récents essaient de modéliser l'information contenue dans les lexiques pour la production automatique de règles ou pour l'apprentissage par analogie (Yvon [Yvo96]).

La variabilité a fait l'objet de travaux spécifiques (Lacheret-Dujour [Lac90]) et commence à être prise en compte dans certains systèmes.

Dans tous les cas, les règles sont structurées sur une dimension, de la plus particulière à la plus générale, même si certains interpréteurs créent automatiquement des structures arborescentes dans un but fonctionnel.

3. Choix d'une structure arborescente dès l'écriture des règles

Lorsqu'on regarde l'évolution des différents phonétiseurs, on constate que le nombre de règles n'a pas cessé de croître, et ceci malgré la percée des lexiques. Les systèmes actuellement opérationnels ont tous dépassé le millier de règles depuis une dizaine d'années. Nous souhaitons montrer ici comment les problèmes aigus de lisibilité et de maintenance posés par l'agencement de ces règles peuvent être sérieusement atténués lorsque l'on envisage une organisation arborescente de ces dernières.

3.1. Introduction d'une nouvelle règle dans une structure à une dimension

Supposons que dans l'état actuel, la règle de rang n s'applique. Un logiciel de traçage peut pointer son emplacement. La nouvelle règle (plus particulière) doit

alors être placée quelque part avant cette dernière. Le choix étant large, on recherche généralement une position pertinente ne dégradant pas la lisibilité. Exemple :

...•an•{cns}...	ɑ̃:	(manche, banque, grange)
(*)...d•am•n...	a:	(damné, condamner)
...•a•mn...	a	(amnistie)
...•a•mm...	a	(gamme)
...•am•{cns}...	ɑ̃:	(ample, chambre)
...•a•...	a	

où : •xxx• = chaîne à transcrire
{cns} = catégorie consonne
... = contexte quelconque

Pour ajouter la règle (*) portant sur *am* dans *damné* et *condamner*, nous trouvons la règle qui s'applique : *a* suivi de *mn* → a *(amnistie)*. Dans notre exemple, la nouvelle règle (contexte *damn*) est placée juste avant.

Si l'introduction d'une nouvelle règle est fiable et relativement rapide quel que soit le nombre de règles, en revanche le choix d'une position pertinente est de plus en plus large et complexe, et par conséquent l'architecture finit par devenir confuse.

3.2. Émergence d'une structure arborescente

Revenons maintenant à notre nouvelle règle : puisqu'elle doit s'appliquer avant la règle n, c'est qu'elle constitue une exception de celle-ci. Il y a donc une connexion implicite entre ces deux règles. L'ensemble des connexions implicites entre règles n'est autre qu'un arbre. Plutôt que de concevoir un interpréteur qui crée cette structure à partir de la liste unidimensionnelle, notre approche consiste à créer nous-mêmes la structure arborescente au moment de l'écriture des règles.

Dans l'exemple que nous avons choisi, la figure 1 illustre les relations entre règles : chaque règle est un noeud de l'arbre, ses exceptions étant les noeuds qui lui sont connectés. Une telle structure optimise le parcours de l'interpréteur en éliminant tous les tests inutiles. Par conséquent elle répond parfaitement aux exigences de vitesse demandées par nos utilisateurs aveugles (réponse instantanée et lecture rapide), nous l'avons donc retenue pour l'écriture de notre phonétiseur Kali, sous la forme décrite en 3.3.

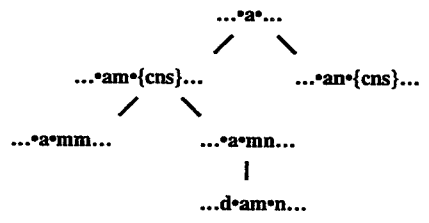


Figure 1. Arbre de dépendance des six règles choisies.

3.3. Introduction d'une nouvelle règle dans une structure arborescente

Si nous reprenons l'exemple de la nouvelle règle à insérer, la méthode est alors la suivante : nous parcourons les règles de niveau 1 (reliées à la racine) jusqu'à trouver celle dont elle dépend (règle mère), puis nous parcourons les règles de niveau 2 reliées à cette dernière, et ainsi de suite, jusqu'à un niveau où aucune règle mère n'est trouvée. Notre règle s'ajoute alors à la liste des règles de ce niveau. Exemple :

I...•a•...	a	
-I...•am•{cns}...	ɑ̃:	(ample, chambre)
--I...•a•mm...	a	(gamme)
--I...•a•mn...	a	(amnistie)
---I...d•am•n...	a:	(damné, condamner)
-I...•an•{cns}...	ɑ̃:	(manche, banque, grange)

Pour ajouter la règle portant sur *am* dans *damné* et *condamner*, nous partons de la racine (*a* → *a*), puis nous trouvons parmi ses exceptions : *am*-consonne → *ɑ̃: (ample, chambre)*. Cette exception comporte elle-même des exceptions (niveau 2), dont : *a* suivi de *mn* → *a (amnistie)*. Notre nouvelle règle étant incluse dans cette dernière, et aucune règle mère n'étant trouvée au niveau 3, c'est là que nous la plaçons, avec le contexte *damn*.

Les exceptions à une règle sont donc placées après celle-ci, avec une indentation supplémentaire (formalisée ici par des tirets) pour marquer leur dépendance. Cette méthode, vieille comme l'écriture, permet de structurer un texte et contient de fait une structure arborescente.

La stratégie de transcription utilisée s'applique de la même façon pour les liaisons, élisions, nombres, sigles et abréviations, qui sont intégrés dans les règles.

3.4. Avantages de la structure arborescente

Dans cette structure, chaque règle a donc sa place, ce qui permet d'éviter la perte progressive de lisibilité. En parcourant l'arbre, nous retrouvons les règles de base du Français près de la racine (*ai*, *ain*, *am*, *an*, etc.), puis dans les niveaux suivants des règles plus particulières (consonnes finales muettes ou non, terminaisons en *ent*, *h* disjonctif ou non, *i*-voyelle, *s* entre voyelles, *ti*-voyelle prononcé *si* ou *ti*, etc.), et aux derniers niveaux les règles les plus particulières.

Des règles qui auraient pu appartenir à des lexiques se trouvent alors structurées. Exemple : l'ensemble des règles sur *...•t•i{voy}...* où *t* est susceptible de se prononcer *s* se décompose en *...{voy}•t•i{voy}...*, *...c•t•i{voy}...*, *...n•t•i{voy}...*, *...p•t•i{voy}...* et *...r•t•i{voy}...*, règles qui elles-mêmes se décomposent plus finement. Nous aboutissons ainsi à une hiérarchisation fine de l'ensemble des règles de transcription.

De plus, comme un arbre s'étend sur deux dimensions, le nombre total de règles est homogène à une surface, tandis que le chemin à parcourir pour atteindre l'emplacement de la règle à insérer est homogène à une distance. Il n'est donc jamais très long.

Plus généralement, la complexité de l'arbre augmente comme la racine carrée du nombre de règles, ce qui permet de gérer sans difficulté, comme nous le verrons plus loin, un nombre de règles très supérieur à 1000.

L'une des conséquences de notre approche est une modification de l'équilibre règles/lexiques, en faveur des règles : le nombre de règles atteint à ce jour est de 4700. Les lexiques (cf. 3.5) ne contiennent actuellement que quelques centaines d'entrées.

3.5. Lexiques

Dans la pratique, les listes de mots ou groupes de mots que nous constituons à partir des erreurs rencontrées sont d'abord placées dans des lexiques, de façon à être immédiatement opérationnels. Mais ce choix n'est pas définitif, et nous les considérons plutôt comme placés en file d'attente. Peu à peu, nous les intégrons dans les règles. Cette procédure nous permet souvent de généraliser : par exemple le mot "Washington" peut être placé dans un lexique. Mais son implémentation dans les règles nous permet de voir que le contexte "ington" suffit et recouvre alors tous les mots en "ington", jusqu'ici oubliés :

---l...•ington• iŋtɔn (Washington, Wellington, etc.)

L'absence du joker "..." signifie que l'extrémité du contexte correspond à l'extrémité du mot.

Le fait que nous cherchions à intégrer les lexiques ne veut pas dire que ceux-ci soient voués à une existence éphémère : nous espérons les développer plus vite que notre capacité à les transformer en règles. Le problème est le suivant : un lexique de noms propres, surtout étrangers, contient beaucoup de mots d'occurrence très rare et doit être étudié avec soin pour éliminer diverses sources d'erreurs, notamment les homographes de mots plus courants. Le rapport efficacité/coût n'est pas forcément supérieur à celui de l'intégration d'une petite liste contenant des mots nettement plus courants, comme on en obtient après un test d'évaluation.

4. Intégration du module de transcription dans le système de synthèse vocale

Bien que le transcripteur puisse fonctionner de façon autonome, ce dernier est destiné à être intégré dans le synthétiseur vocal Kali. Il bénéficie par conséquent des informations fournies par une analyse syntaxique automatique, qui, outre un étiquetage du texte à synthétiser en catégories morphosyntaxiques, propose, sur les bases de règles de dépendance (Giguet et Vergne [Gig97]), une segmentation du texte en constituants intonosyntaxiques.

L'utilisation des résultats fournis par l'analyseur syntaxique en entrée de la conversion graphème-phonème permet un traitement fin de la liaison ainsi qu'une bonne résolution des problèmes posés par la transcription des homographes hétérophones. C'est ainsi qu'une liaison *a priori* associée à la finale consonantique d'un adjectif est bloquée par l'adjonction d'un marqueur pour tout adjectif postposé (*i.e.* final de constituant intonosyntaxique) :

un ancien administrateur	liaison obligatoire
ce meuble ancien attire le regard	liaison interdite

En ce qui concerne les homographes hétérophones, si l'absence de traitement spécifique donne le bon résultat dans 80 % des cas environ (ex : *président* = nom, *tous* = pronom), l'analyse syntaxique permet d'atteindre un taux de transcription correcte de 90 % (ex : *Jean et Jacques président, ils viennent tous / ils viennent tous les jours*) ; la résolution des ambiguïtés restantes nécessitant une analyse sémantique poussée non disponible à ce jour.

La mise en œuvre de l'ensemble du synthétiseur vocal pour les essais de transcription permet également de mettre à l'épreuve certains choix qui demandent une confirmation ergonomique. Par exemple le traitement des liaisons et élisions facultatives fait l'objet de tests d'écoute auprès de nos utilisateurs aveugles : l'option choisie doit contribuer à l'esthétique du texte énoncé sans dégrader son intelligibilité.

5. Maintenance des règles

5.1. Écriture et compilation

Une fois les règles correctement positionnées, écrites et agrémentées d'exemples, un compilateur vérifie la syntaxe de ces règles avant de les condenser dans un fichier destiné à l'interpréteur et d'en extraire un fichier de test ne conservant que les exemples. Un comparateur fournit ensuite une liste différentielle d'exemples entre la nouvelle version et la version précédente. Cette méthode garantit le maintien des règles considérées comme acquises : si les nouvelles règles ont été correctement implémentées, seuls les exemples correspondant à celles-ci apparaissent, dans le cas contraire, les exemples erronés pointés par le comparateur renseignent sur les corrections à effectuer.

5.2. Mise à jour

Le fichier de règles évolue en s'enrichissant d'exemples trouvés dans les ouvrages spécialisés et de listes fournies par les utilisateurs de synthèses vocales. La constitution de ces listes à partir de la lecture de journaux et revues augmente la probabilité de traiter les règles dans le bon ordre de fréquence (donc d'importance ?), car ces publications recouvrent des domaines très divers. Le traitement des noms propres d'actualité est indispensable à la bonne lecture de ces textes. Une mise à jour périodique est nécessaire, mais peu à peu, l'ensemble se consolide.

Cette variété permet également de dresser un bilan fiable du niveau de performance, dont l'évolution pourra renseigner sur la pertinence des principes mis en place.

6. Performances actuelles

Nous distinguons deux types de performances, le taux de réussite et la vitesse. Pour être acceptable, ou mieux, pour être apprécié des utilisateurs, un système de synthèse vocale doit réunir ces deux qualités : réponse immédiate et grande majorité des phrases correctes.

6.1. Vitesse de transcription

Les temps de réponse du phonétiseur sont ramenés à ceux qui seraient obtenus sur Pentium 100 MHz, considéré comme configuration minimale.

Le résultat sans analyse syntaxique est de 100 µs par caractère en moyenne. L'analyse syntaxique, plus lente, porte le délai à environ 700 µs par caractère, soit un délai total de **70 ms** pour une phrase de **100 caractères**. Ce temps de réponse presque imperceptible est jugé satisfaisant par les utilisateurs.

6.2. Taux de bonne transcription

La dernière évaluation (avril 1998) de notre phonétiseur repose sur la transcription d'échantillons du journal "Le Monde" en date du 30 mars 1998. Seize articles couvrant les diverses rubriques ont été phonétisés (total 115657 caractères soit 21000 mots soit 77000 phonèmes). Le résultat global est de **99,6 %** de phonèmes corrects. La liste ci-dessous montre le nombre et le taux relatif d'erreurs par catégorie :

Noms propres	150	49 %
Mots d'emprunt	30	10 %
Sigles	11	3 %
Liaisons obligatoires	23	8 %
Liaisons interdites	12	4 %
Élisions obligatoires	11	3 %
Élisions interdites	2	1 %
Homographes hétérophones	32	11 %
Divers	35	11 %
Total	306	100 %

Erreurs potentielles évitées grâce à l'analyse syntaxique	39	13 %

7. Conclusion

Nous avons mis au point une stratégie de transcription et un formalisme qui nous semblent cohérents et qui permettent d'assurer une progression soutenue des performances.

La construction manuelle de la structure arborescente telle que nous la concevons aujourd'hui favorise la maintenance et l'évolution du système. Cette structure offre en outre une lisibilité optimale des règles : elle met en évidence des regroupements de règles selon des critères allant des plus généraux aux plus particuliers, tout en respectant la variété des règles et leur hiérarchie.

Afin d'optimiser notre système, un programme d'automatisation de la structure a été amorcé, comportant notamment : adjonction automatique d'une règle, traçage automatique des règles, historique sur l'utilisation de chaque règle.

Bibliographie

[Aub85] Aubergé V. (1985), "Contribution à la phonétisation automatique des langues alphabétiques : le langage TOPH", Rapport de DEA, CRISS, Université de Grenoble II.

[Bai92] Bailly G., Alissali M. (1992), "Compost : a server for multilingual text-to-speech system", Traitement du Signal, vol. 9, pp. 359-366.

[Bel92] Belrhali R., Libert L., Aubergé V., Boë L.J. (1992), "Des lexiques aux règles : vers une méthode descriptive de la phonétisation du Français", 19èmes JEP, Bruxelles, pp. 225-230.

[Bel95] Belrhali R. (1995), "Phonétisation automatique d'un lexique général du Français : systémique et émergence linguistique", Thèse de l'Université Stendhal de Grenoble 3.

[Bou97] Boula de Mareuil P. (1997), "Étude linguistique appliquée à la synthèse de la parole à partir du texte", Thèse de l'Université de Paris XI, Orsay.

[Cat84] Catach N. (1984), "La phonétisation automatique du Français", Éditions CNRS, Paris.

[Div77] Divay M., Guyomard M. (1977), "Conception et réalisation sur ordinateur d'un programme de trancription graphèmo-phonétique du Français", Thèse de troisième cycle, Université de Rennes.

[Div84] Divay M. (1984), "De l'écrit vers l'oral ou contribution à l'étude des traitements des textes écrits en vue de leur prononciation sur synthétiseur de parole", Thèse d'Etat, Université de Rennes.

[Gig97] Giguet E., Vergne J. (1997), "From part of speech tagging to memory-based deep syntactic analysis", International Workshop on Parsing Technologies 1997 Proceedings, Boston.

[Lac90] Lacheret-Dujour A. (1990), "Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteurs", Thèse de l'Université de Paris VII.

[Lap88] Laporte É. (1988), "Méthodes algorithmiques et lexicales de phonétisation de textes, applications au Français", Thèse de troisième cycle, Université de Paris VII.

[Lec72] Le Cornec A. (1972), "La transcription orthographique-phonétique du Français". Essai de formalisation, Rapport de stage, CNET Lannion.

[Lét80] Léty M. (1980), "Transcription orthographique-phonétique : un système interpréteur", Thèse de doctorat, Université Scientifique et Médicale de Grenoble.

[Pro80] Prouts B. (1980), "Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur", Thèse de Docteur-Ingénieur, Université de Paris Sud-Orsay.

[Tih91] Tihoni J., Pérennou G. (1991), "Phonotypical transcription through the GEPH expert system", Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), pp. 767-770.

[Yvo96] Yvon F. (1996), "Prononcer par analogie : motivation, formalisation et évaluation", Thèse de l'ENST.

CONTOUR INTONATIF ET SYNTAXE EN ARABE.RESULTATS PRELIMINAIRES.

Z. NAJIM

ENSERB, Université de BordeauxI
B.P. 99 Talence Cédex.
Tel: + 33 5 56 84 61 85, Fax: + 33 5 56 84 84 06
Email: najim@goelette.tsi.u-bordeaux.fr

ABSTRACT

To what degree the peaks of the fundamental frequency curve, in Arabic speech, are correlated to the syntactic structure of the read text? The Intonative Model of Rabat (MIR) based on the hypothesis of a correlation is actually integrated in an Arabic text-to-speech system. This paper aims to present a critical look at the MIR in order to improve the naturalness of the synthetic speech. The present analysis is limited to the verbal and nominal affirmative phrases. It is based on the approach developed at the Institute Phonétique d'Aix, particularly using the system INTSINT for prosodic coding.

1. INTRODUCTION

Le rapport entre la prosodie et la syntaxe, a été longuement débattu par les linguistes, [Mar80a et b], [Sel84] et [Hir]. Pour certains, l'intonation d'un énoncé n'est pas déterminée directement par la structure syntaxique, mais par une structure phonologique qui est un médiateur entre la syntaxe et la prosodie. Pour d'autres, par contre, c'est la structure syntaxique qui détermine directement la prosodie. En ce qui concerne l'arabe littéral, c'est plutôt cette position qui a été adoptée. Il est vrai que nous ne disposons que de peu de travaux concernant l'analyse instrumentale des faits prosodiques en Arabe (voir pour l'arabe marocain [Ben82]. Un regain d'intérêt pour ce domaine a accompagné le développement des recherches concernant le traitement automatique de la parole arabe et en particulier la synthèse à partir du texte. Dans ce cadre, les travaux de [Raj89] et [Esk88] portent un intérêt particulier aux aspects rythmiques et intonatifs de l'arabe moderne dans le but de développer et d'implémenter un modèle prosodique ; le naturel de la parole synthétisée dépend intimement des performances d'un tel modèle.

Par ailleurs un problème particulièrement intéressant au niveau méthodologique se pose lors de toute étude de l'arabe parlé, à savoir le problème du locuteur natif. En effet, il n'existe point de référence orale pour la langue arabe, le texte coranique accepté comme étant la référence sacrée pour la définition de l'arabe, n'a été transmise que sous sa forme écrite. L'étendue géographique du monde arabophone a contribué à l'émergence de fait de parlers arabes loin d'être identiques.

2. DESCRIPTION ET IMPLEMENTATION DU MIR.

2.1. Description du modèle.

Le calcul du contour intonatif d'une phrase déclarative est basé sur les faits suivants :
- la place de l'accent lexical est sauvegardée dans la phrase,
- les maxima intonatifs sont placés sur les syllabes accentuées des mots de la phrase,
- le contour intonatif de la phrase est constituée de contours primitifs correspondant chacun à un mot,
- un contour primitif est constitué de deux segments linéaires montant et descendant,
- le maxima du contour primitif dépend de la structure syntaxique de la phrase porteuse et de l'attribut syntaxique propre au groupe syntaxique contenant le mot,
- la valeur d'attaque au début de la phrase est 120 Hz, la valeur finale est 100 Hz.
- les informations syntaxiques sont déterminées à partir d'un certain nombre de marqueurs intonatifs positionnés manuellement par l'opérateur dans la phrase à traiter.

2.2. Implémentation.

Les marqueurs exploités pour la synthèse de l'intonation de la phrase déclarative sont :
(£) : marqueur placé après le verbe dans une phrase verbale. Il signifie une montée intonative de +20Hz.
(*) : marqueur placé après le groupe nominal sujet (GN1) d'une phrase nominale ou verbale. Il signifie une montée intonative de +25Hz sur chaque mot du groupe sujet du verbe et une montée de +30Hz dans le cas de la phrase nominale.
(&) : marqueur placé après le groupe complément GN2 d'une phrase verbale. Il indique une montée intonative de +25Hz sur chaque mot du groupe complément du verbe.
($) : marqueur placé après le groupe nominal attribut GH2 de la phrase nominale. Il signifie une montée intonative de +35Hz sur chaque mot du groupe GN2.
(%) : marqueur de fin de phrase. Il indique une descente intonative finale de -20Hz par rapport à la dynamique de base (120Hz).

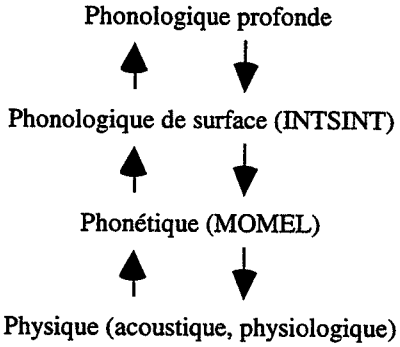
Le traitement intonatif est réalisé à partir d'un positionnement manuel de marqueurs intonatifs sur la chaîne phonétique sans exploitation des signes de ponctuation. Ces marqueurs sont introduits en des points de l'énoncé considérés comme pertinents de vue syntaxique.

La sortie est constituée par la courbe intonative décrivant les variations des valeurs de la fréquence fondamentale de la phrase en fonction du temps. Cette courbe est constituée d'autant de primitives que de mots contenus dans la séquence, chaque primitive étant constituée de deux pentes montante-déscendante. La première commence au début du mot à 120Hz (dynamique de base) et se termine aux 2/3 de la voyelle accentuée à une valeur de F0 déterminée par le type du marqueur suivant le mot. La pente descendante est adjacente à la pente ascendante et finit à la fin du mot à 120Hz pour un mot en position non finale. Dans le cas du mot final, elle descend à 100Hz.

3. LE MODELE AIXOIS: PRINCIPE ET OUTILS.

3.1. Principe.

Le modèle aixois décrit dans [Hir87] et [Hir] propose une formulation universalisante de la représentation phonologique de la prosodie. Le modèle fait appel à :
- quatre niveaux de représentation :



- et deux principes ; le principe de non spécifité qui est une théorie complète où l'on décrit l'intonation avec les mêmes éléments phonologiques que ceux utilisés par les systèmes lexicaux (i.e. ton, accent,...). Le principe d'interprétabilité qui permet de définir les différents niveaux en termes sous jacents. L'intonation est l'interface entre la phonologie et l'acoustique.

3.2 Système de transcription INTSINT.

Le système INTSINT (International Transcription System for Intonation) permet de décrire les différents points cibles de la courbe de fréquence fondamentale (F0) par deux classes de symboles qui spécifient la hauteur des cibles. Ces symboles sont soit absolus (Mid, Top, Bottom), soit relatifs se référant à la valeur fréquentielle précédente (High, Low, Down, Hop, Same).
Ce système développé à Aix a été particulièrement utilisé pour les études de l'intonation du Français, de l'Espagnol et de l'Anglais [Hir93].

4. ANALYSE ET EVALUATION DU MIR.

Parmi les finalités de nos études sur l'intonation de l'Arabe [Naj93, 94, 95] nous nous sommes particulièrement intéressés à l'évaluation du MIR à partir du modèle aixois. Le corpus analysé et codé par INTSINT est constitué de 19 phrases déclaratives, nominales et verbales, lues par un locuteur marocain. Ces phrases sont constituées de deux à six mots à partir d'un lexique de 14 mots englobant différentes structures syllabiques et schémas accentuels (voir annexe 1). Après étiquetage manuel et découpage syllabique, l'analyse et le codage ont permis le calcul des contours intonatifs (fonction peigne), le codage des points cibles et la localisation des sommets intonatifs. Des corrections manuelles basées sur un contrôle auditif ont été nécessaires pour corriger les points cibles mal détectés. Nous donnons ci-dessous des exemples des résultats obtenus (les chiffres correspondent aux valeurs de F0 des points cibles, en dessous sont notées les valeurs absolues et relatives selon INTSINT) :

```
 kataba
163   149   110  96
M     D     D    B

     kataba          lwaladu
145   161  148      157    82
M     T    L        H      B

      kataba       lwaladu      lkabiiru
156   162  161     150          148  90
M     T    D       D            D    B

   kataba      lwaladu  lkabiiru   rissaalatan
145   156   159  153  157   155   137 146 83
M     U     T    L    H     D     L   H   B
```

La description fournie par MIR donne des valeurs brutes en Hz, avec des variations de -20 et +35Hz. Les courbes intonatives ainsi obtenues sont une succession de pics (maxima) et de creux sans aucun effet de déclinaison.

Tableau 1: Analyse des valeurs des maxima intonatifs des deux sous corpus. (N est le nombre des maxima)

Corpus	N	Moy	Ecart type
Phrases verbales	30	154,5	10,7
Phrases nominales	19	155	4,2

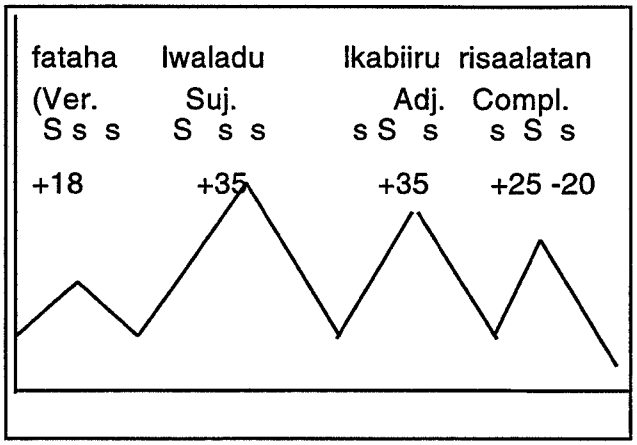


Figure 1: profil des contours intonatifs obtenus par MIR, adapté à l'une de nos courbes.

Dans la définition des marqueurs, le MIR montre que les pics sont associés à la structure syntaxique, et qu'ils sont plus élevés sur les phrases nominales que sur les phrases verbales. Notre analyse des deux sous corpus constitués respectivement par les phrases verbales et les phrases nominales a donné les résultats présentés dans le tableau1.

L'analyse de variance, ne donne aucune différence significative, F(1,47) = 0,052, p= 0,8212.Ces résultats infirment l'hypothèse du MIR relative à l'influence du type de phrase sur les hauteurs des pics intonatifs.
Par ailleurs, dans les règles qui déterminent la hauteur des pics, le MIR fait intervenir les attributs syntaxiques de chaque mot de la phrase : les compléments se situant à un niveau plus élevé que les verbes, lesquels sont à un niveau plus élevé que les sujets. Nous avons cherché à évaluer sur notre corpus l'influence de l'attribut syntaxique sur la hauteur des pics. Les résultats obtenus sont représentés dans le tableau ci-dessous.

Tableau 2: Valeurs des pics intonatifs en fonction de la nature syntaxique des constituants.

corpus	N	Moy	Ecart type
Verbes	11	166,4	5,4
Sujets	12	162,8	4,7
Complément	28	154,7	6,8

L'analyse de variance donne une différence hautement significative, F (2,48) =17,516 p=0,0001, ce qui semble conforter les résultats du MIR sur l'influence de la nature syntaxique. Cependant la moyenne pour les compléments (154,7 Hz), est plus basse que celle des verbes (166,4 Hz) et des sujets, (162,8 Hz). Ainsi, contrairement à l'ordre fourni par MIR (le verbe est plus haut que le sujet, plus haut que le complément), la hiérarchie observée dans notre analyse, est plutôt que le verbe est plus haut que le sujet qui est plus haut que le complément. L'explication que nous proposons, est que cette hiérarchie, correspond à l'ordre linéaire des phrases du corpus. Plus un constituant se situe vers la fin d'un énoncé, plus son maxima intontif est moins haut. Or, le modèle intonatif de Rabat, ne donne aucun effet de déclinaison.

D'autre part le système de transcription INTSINT montre l'existence de variations locales d'ordre micromélodique que le schéma squelettique du MIR ne permet pas de restituer.

Conclusion et perspectives :

Les résultats présentés dans ce travail, pour l'intonation des phrases déclaratives permettent d'affirmer les faits intonatifs suivants :
- Les maxima des contours intonatifs seraient corrélés à la structure syntaxique.
- Les schémas des variations données par le MIR devraient être corrigés dans le sens où le verbe est plus haut que le sujet qui est plus haut que le complément ;
- L'introduction d'un effet de déclinaison serait nécessaire.

Des corrections potentielles sont actuellement en cours de validation par un système de synthèse de l'arabe, à partir du texte. Ce travail est actuellement étendu aux différents types de phrases.

Bibliographie:

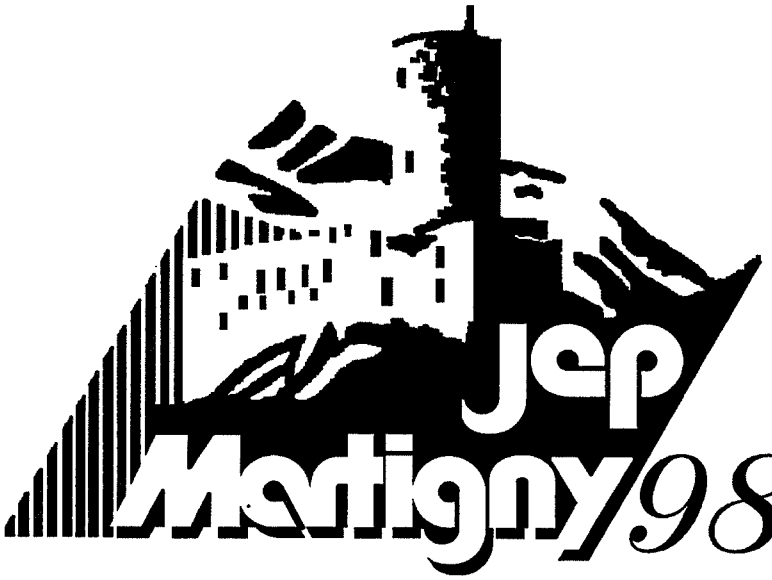
[Mar80a] Martin Ph. (1980), " Sur la non congruence entre les structures syntaxiques et prosodiques", TIPA, 7, pp: 319-339.
[Mar80b] Martin Ph. (1980), "L'intonation est-elle une structure congruente à la syntaxe?", in: Rossi et al. L'intonation de L'acoustique à la sémantique, Kliencksieck, 234-271.
[Sel84] Selkirk E.O. (1984), Prosody & Syntax: The realtion between sound and structure. M.I.T press, Cambridge.
[Esk87] Es-skalli L. et al. (1987), Eléments d'un modèle intonatif pour la synthèse de la parole arabe , Actes 15èmes JEP, pp: 282-285.
[Hir87] Hirst D.J. (1987), La description linguistique des systèmes prosodiques une approche cognitive. Thèse de doctorat d'état. Université de Provence.
[Raj89] Rajounai A. et al. (1989), "Etude de l'accent lexical en Arabe", Journal d'Acoustique,2, pp: 171- 176.
[Hir93] Hirst D.J. et al. (1993), Multi-lingual modelling of intonation patterns. Proc. ESCA Workshop on Prosody, Lund University.
[Naj93] Najim Z. (1993), Prosodie de l'arabe littéral parlé au Maroc: localisation des sommets mélodiques. Actes Colloque 'Phonologie, phonétique et prosodie', Toronto.
[Naj94] Najim Z. (1994), Codage prosodique d'un corpus d'arabe littéral lu par un locuteur marocain. TIPA, 15, pp:87-98.
[Naj95] Najim Z. (1995), Prosodie de l'arabe standard parlé au Maroc, analyse historique, sociolinguistique et expérimentale. Thèse de Doctorat, Université de Provence.

[Naj97] Najim Z. et al. (1997), A perceptual study of Arabic intonation . Proc. Intonation : Theory, Models and Applications, ESCA Workshop, pp: 279-283
[Ben] Benkirane T. Intonation in Westen Arabic (Moroccan). In Hirst, D. et Di Cristo, A (à paraître). (eds).
[Hir] Hirst D.J.& Di Cristo A. (eds) (à paraître), Intonation Systems: a Survey of Twenty Languages. Cambridge University Press.

Annexe 1.

1 kataba	(il a lu)
2 alwaladu	(le garçon)
3 alkabiiru	(le grand)
4 risaalatan	(une lettre)
5 'akala	(il a mangé)
6 mousa	(Moussa)
7 filmadrasati	(à l'école)
8 fataha	(il a ouvert)
9 baaba	(la porte)
10 fii ssabaahi	(le matin)
11 al bataatesa	(les pommes de terre)
12 ttuffaahata	(une pomme)
13 fii ddaari	(à la maison)
14 almatbuuxa	(cuite)

Parole et émotion



Perception de l'émotion : les indices prosodiques et la durée minimale

Soo-Jin Chung*
Institut de Phonétique, Université de la Sorbonne Nouvelle, Paris, France.

Abstract

The present study reports two experiments on the perception of emotion in a real situation. The results showed that (i) the emotional upset of speaker was better perceived from the final part of the utterance than from the other parts, and that (ii) listeners needed at least 100ms, the duration of a vowel, to perceive the speakers' emotion. The perception of emotion was determined by pitch level and jitter, depending on the type of stimuli.

Introduction

L'état émotionnel s'exprime par des changements prosodiques de la voix par rapport à la voix neutre et l'auditeur s'appuie sur ces changements pour percevoir l'émotion du locuteur, ce qui est appelé 'utilisation des indices acoustiques' ([Sch77]). Beaucoup de chercheurs se sont intéressés aux problèmes de la communication de l'émotion (voir [Sch86]) mais peu des résultats sont comparables à cause des différences méthodologiques. Les différences portent sur les émotions étudiées, sur l'unité d'analyse de la parole émotionnelle (syllabe, mot, phrase) et la portée de la description n'est pas souvent clairement établie. Est-ce que la perception de l'émotion est stable quelle que soit la taille et la position de la fenêtre d'analyse dans la phrase? La première expérience de cette étude s'intéresse à cette question et compare des stimuli qui correspondent à une phrase entière et à des parties de la phrase. Concernant la minimalité de la perception de l'émotion, la plupart des études ont été faites dans le domaine fréquentiel (par ex. [Sch72]) mais peu dans le domaine temporel. La deuxième expérience de cette étude vise à pallier ce manque, et examine la perception d'une émotion forte selon la taille du stimuli dans le domaine temporel.

Expérience I

L'expérience I examine si l'émotion du locuteur est mieux exprimée et mieux perçue dans la partie finale de la phrase que les autres. L'hypothèse de la primauté de la partie finale de la phrase vient du fait suivant. On sait que la prosodie a des fonctions linguistiques. Elle peut marquer la modalité de la phrase, par exemple. La phrase affirmative et la phrase interrogative peuvent se distinguer par l'intonation finale, et l'allongement final signale la fin de phrase ([Vai83]). Dans le cas des fonctions paralinguistiques de la prosodie, il est noté que le cliché mélodique de la profession ou de l'attitude est souvent manifesté dans la partie finale de la phrase ([Fon83] et [Leo93]). Quand à l'expression des émotions qui n'est pas toujours sous le contrôle du locuteur, est-elle également localisée dans la partie finale, comme les deux fonctions précédentes?

1.1. Corpus

Le corpus a été établi à partir de l'enregistrement de l'interview télévisé. Cinq locutrices Américaines, âgée de trente ans environ, fondaient en larmes en racontant leurs expériences douloureuses. Elle n'étaient ni actrices ni personnes connues du public et il n'y avait aucune tentative de déguiser, de supprimer ou d'exagérer l'émotion. Leur expression de l'émotion peut être considérée authentique et naturelle, différente de l'émotion imitée ou stylisée. Les énoncés neutres étaient extraits du moment où les locutrices parlaient sans émotion particulière, et les énoncés émotionnels au moment où elles parlaient 'sous le coup de l'émotion de détresse'.
On a préparé deux types de stimuli pour le test de perception, 'stimuli de phrase (Phrases)' et 'stimuli de partie (Parties)'. Les Phrases correspondaient à 48 phrases entières, 24 neutres et 24 émotionnelles. Les Parties étaient le résultat de la segmentation des Phrases en trois parties, initiale, médiane et finale, soit 144 parties (48x3 stimuli).

1.2. Test auditif

Vingt-deux auditeurs Américains, étudiants de différents départements à l'université de Brown ont participé au test de perception. Leur âge variait de vingt-et-un ans à trente ans. Les auditeurs ont été divisés en deux groupes. Chaque auditeur a écouté un des deux types de stimuli d'une locutrice donnée, pour éviter tout effet d'apprentissage. Les stimuli neutres et émotionnels étaient présentés une fois, dans un ordre arbitraire et les auditeurs devaient juger si la locutrice pleurait ou pas pendant deux secondes d'intervalle. Le test total a duré moins d'une heure.

1.3. Analyses acoustiques

La moyenne de la fréquence fondamentale (le F0 moyen) et la durée de chaque stimulus étaient mesurées à l'aide des programmes d'analyse du signal 'Waved' et 'Pitch' dans le laboratoire de phonétique de l'université de Brown. Les programmes ont été développés par John Mertus dans le laboratoire. Ils permettent la visualisation du signal de la parole, l'édition du signal, l'analyse spectrographique, et l'estimation de la variation de F0.

* Adresse acturelle: Cognitive and Linguistic Science Department, Brown University, Providence, RI 02906, USA.

1.4. Evaluation statistique

Les analyses de variance ont estimé l'effet du type de stimuli et de la position de la partie, sur la perception de l'émotion. La variable dépendante était la réponse des auditeurs. La réponse correcte était déterminée par la correspondance entre la catégorie des stimuli dans le corpus, émotionnel ou neutre, et la réponse des auditeurs. Les variables indépendantes étaient les facteurs du 'type de stimuli' et de la 'position de la partie dans la phrase,' initiale, médiane ou finale. L'effet de l'émotion sur les paramètres prosodiques a été testé par le t-test. La relation entre le 'taux de la réponse correcte' et les paramètres acoustiques a été estimée par l'analyse du coefficient de corrélation de Pearson.

1.5. Résultat

Type de stimuli L'émotion a été significativement mieux reconnue dans les Phrases que dans les Parties ($F(1,2011)=38.3$, $p<0.01$). Le taux de la réponse correcte passe de 85.8% à 75.4% des Phrases aux Parties. Cet effet du type de stimuli n'est pas surprenant mais il n'était jamais été testé auparavant d'une manière systématique.

Table 1. Taux (%) de la réponse correcte.

Stimuli	Phrase	Partie	Total
Locutrice 1	83.3	77.5	80.4
Locutrice 2	84.1	70.0	77.1
Locutrice 3	80.7	71.6	76.2
Locutrice 4	84.1	67.4	75.8
Locutrice 5	98.9	92.8	95.9

Position de la partie dans la phrase La perception de l'émotion était influencée par la position de la partie dans la phrase ($F(2,767)=1.76$, $p<0.05$). D'après l'analyse post-hoc, le taux de réponses correctes correspondant à la position finale est significativement plus élevé que celui des autres positions, initiale et médiane, et il est presque égal au taux des Phrases entières (Figure 1). Ce résultat indique que l'émotion du locuteur est mieux reconnu à partir de la partie finale que dans les autres parties.

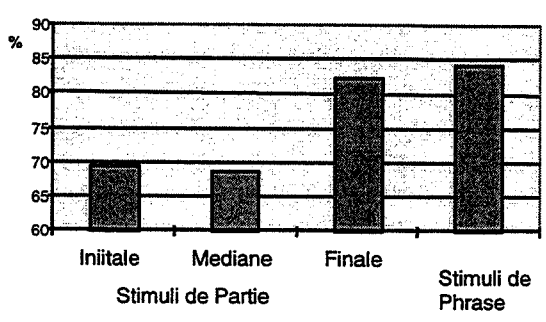


Figure 1. Taux moyen des réponses correctes.

Durée La durée des Phrases varie de 650ms à 1750ms et celle des Parties de 150ms à 360ms. La durée moyenne des Phrases et des Parties est de 1107.7ms et 247.5ms, respectivement. La différence de durée moyenne entre les Phrases et les Parties semble avoir influencé la perception de l'émotion. Cependant, pour un type de stimuli donné, les Phrases ou les Parties, la longueur de la durée n'est pas corrélée au taux de réponses correctes. Autrement dit, l'émotion des stimuli plus longs n'était pas nécessairement mieux reconnue que celle des stimuli plus courts. Il faut noter que la différence de durée est plus grande entre la phrase la plus longue et la phrase la plus courte qu'entre la durée moyenne des Phases et celle des Parties. Mais le taux de réponses correctes de la phrase la plus longue n'était pas très différent de celui de la phrase la plus courte.

F0 moyen Quand les locutrices parlaient en pleurs, le F0 de leur voix était plus élevé, d'où le F0 moyen des stimuli émotionnels plus élevé que celui des stimuli neutres ($t(190)=-6.12$, $p<0.01$). L'augmentation du F0 moyen a varié largement selon les locutrices. Les réponses des auditeurs sont cependant corrélées avec la valeur du F0 moyen ($r(97)=0.43$, $p<0.01$).

Table 2. F0 moyen (Hz) de cinq locutrices

Stimuli	Neutres	Émotionnels
Locutrice 1	255	320
Locutrice 2	212	214
Locutrice 3	207	228
Locutrice 4	195	251
Locutrice 5	192	305
Total	216.7	264.0

Débit L'émotion a ralenti le débit de parole des locutrices, donc le débit était plus lent pour les Phrases émotionnelles que les Phrases neutres. Mais ce ralentissement n'était pas statistiquement significatif ($F(1,46)=0.69$, $p>0.05$) et le débit était à peine corrélé au taux de la réponse correcte ($r(264)=-0.23$, $p=0.048$).

En résumé, le résultat de l'expérience I montre que l'émotion est mieux perçue dans la partie finale que les autres parties de la phrase et que F0 moyen est un bon indice prosodique pour la perception de l'émotion de détresse. En termes acoustiques, F0 moyen s'est élevé significativement en cas d'émotion et le débit s'est ralenti un peu. Malgré la grande variation du F0 moyen des locutrices, les auditeurs étaient capables de reconnaître ces changements prosodiques dûs à l'émotion. Leur perception correcte des stimuli était corrélée avec les changements typiques des paramètres acoustiques. Ce résultat confirme la proposition de Scherer et Oshinsky [Sch77], à avoir que l'auditeur perçoit les changements acoustiques de la voix émotionnelle par rapport à la voix neutre et les utilise comme indices de l'émotion dans le test de perception. Le taux élevé de la perception de l'émotion malgré la grande variation des locutrices dans cette expérience

est aussi en accord avec l'idée de Streeter et al. [STR83] que les corrélats perceptifs de l'émotion sont plus stables que les indices acoustiques de l'émotion.

2. Expérience II

Dans l'expérience I, la perception de l'émotion était sensible au type des stimuli, phrase entière vs. partie, plutôt qu'à la différence de la durée absolue des stimuli. D'autre part, l'émotion pouvait être correctement perçue même dans quelques stimuli très courts. Cependant, on peut bien imaginer qu'il y ait une durée minimale dont l'auditeur a besoin pour percevoir l'émotion du locuteur. D'après la littérature, l'émotion peut être bien transmise dans un seul mot ou une voyelle (Kaiser [Kai62] et Leinonen et al. [Lei97]). L'expérience II s'intéresse à la durée minimale de la perception de l'émotion pour savoir dans quelle mesure l'auditeur peut percevoir correctement l'émotion de locuteur.

2.1. Corpus

Trois types de stimuli avec les durées différentes, 150ms, 100ms et 50ms, ont été établis à partir des stimuli de parties (Parties) utilisés dans l'expérience précédente. On a sélectionné vingt-quatre Parties dont l'absence ou la présence de l'émotion était parfaitement reconnue dans le test de perception précédent, pour deux locutrices, locutrice 1 et 5. Les stimuli sélectionnés étaient pour moitié neutres et pour moitié émotionnels. Dans les Parties originales dont la durée variait de ??? à 280ms, la durée moyenne des voyelles était de 102.3ms. Afin d'établir les trois types de stimuli, on a extrait 150ms à partir du centre d'une voyelle de chaque partie et a réduit sa durée à 100ms, puis à 50ms. Ainsi, les '50ms-stimuli' contenaient moins de 50ms et les '150ms-stimuli' plus de 50ms que la durée moyenne des voyelles.

2.2. Test auditif

Treize auditeurs Américains qui n'avaient pas participé à l'expérience précédente ont passé le test de perception. Leur âge était de 20 ans à 28 ans. La tâche et la procédure du test étaient le même que pour le test précédent sauf une seconde de la durée intervale entre les stimuli. L'ordre de la présentation des trois types de stimuli était contrebalancé pour prévenir tout apprentissage. Le test total a duré vingt minutes.

2.3. Analyses acoustiques

Les paramètres acoustiques, le F0 moyen et la perturbation de F0 ('jitter'), ont été mesurés avec les mêmes programmes que ceux utilisés dans Expérience I. La perturbation de F0 est estimée par la moyenne des différences de fréquence des périodes adjacentes dans un stimulus donné.

2.4. Evaluation statistique

L'analyse de variance a examiné l'effet de réduction de la durée dans les stimuli minimaux. La variable dépendante était la réponse des auditeurs et la variable indépendante du 'type de stimuli' avait trois niveaux, 50, 100 et 150ms.
La différence des indices acoustiques, le 'F0 moyen' et la 'perturbation de F0,' entre les stimuli émotionnels et les stimuli neutres est estimée par le t-test. La corrélation entre les indices acoustiques et la perception de l'émotion est évaluée par l'analyse de corrélation.

2.5. Résultat

Type de stimuli La réduction de durée a affecté la perception de l'émotion ($F(2,933)=8.7$, $p<0.01$). Plus les stimuli étaient courts, plus les auditeurs ont fait d'erreurs dans la perception de l'émotion. Cet effet n'est cependant significatif qu'entre les '100ms-stimuli' et les '50ms-stimuli,' mais pas entre les '150ms-stimuli' et les '100ms-stimuli'. Les taux de la réponse correcte des 150ms-stimuli et des '100ms-stimuli' montrent un effet de plafonnement et on pourrait situer le seuil d'intégration à la durée de 100ms.

Table 3. Taux de la réponse correcte (%).

Stimuli minimaux	Neutres	Émotion	Total
150ms-stimuli	94.9	86.5	90.7
100ms-stimuli	92.3	80.8	86.5
50ms-stimuli	87.9	70.5	79.2

F0 Moyen Les stimuli émotionnels avaient un F0 moyen plus élevé que les stimuli neutres ($t(70)=-7.29$, $p<0.01$), comme prévu. Cependant, contrairement au résultat d'Expérience I, la valeur du F0 moyen n'était pas systématiquement corrélée au taux de perception de l'émotion dans les stimuli minimaux. De plus, l'émotion était même mieux percue dans les stimuli avec un F0 moyen relativement bas que dans les stimuli avec un F0 moyen extrêmement haut.

Table 4. F0 moyen (Hz) des stimuli des deux locutrices.

	Neutres	Émotion	Moyen
Locutrice A	248.3	363.3	305.8
Locutrice B	183.3	270.8	227.1

Perturbation de F0 ('jitter') Le bouleversement émotionnel des locutrices a augmenté la perturbation de F0 ('jitter'). La perturbation de F0 était plus grande pour les stimuli émotionnels que les stimuli neutres ($t(70)=-10.85$, $p<0.01$). Cette différence était reflétée dans la perception de l'émotion. C'est-à-dire, les stimuli émotionnels avec une perturbation de F0 plus grande étaient mieux reconnus que ceux avec une moindre perturbation de F0 ($r(234)=0.18$, $p<0.01$).

Table 5. Perturbation de F0 des stimuli minimaux.

	Neutres	Émotion	Moyen
Locutrice A	0.034	0.062	0.048
Locutrice B	0.025	0.068	0.047

En résumé, le résultat d'Expérience II montre que la perception de l'émotion est significativement dégradée quand l'auditeur écoute un énoncé de moins de 100ms, la durée moyenne d'une voyelle dans cette expérience. Concernant l'indice prosodique, les valeurs du F0 moyen et de la perturbation de F0 des stimuli émotionels sont plus haut que celles des stimuli neutres. L'augmentation de la perturbation de F0 confirme la prédiction de Scherer [Sch86] mais n'est pas conforme au résultat de Protopapas et Liberman [Pro97]. La différence méthodologique de l'estimation de 'jitter' et de l'émotion étudiée semble contribuer à ce désaccord. A la différence du résultat de l'Expérience I, la perturbation de F0 a signalé la présence de l'émotion dans les stimuli minimaux plutôt que le F0 moyen, ce qui suggère une autre stratégie possible de l'auditeur dans la perception de l'émotion.

3. Conclusion et Discussion

L'expression et la perception de l'émotion ont souvent été étudiées dans un laboratoire avec les phrases standardisées qui étaient exprimées par les acteurs. Ce genre d'étude sur l'émotion à partir d'un corpus d'émotions stylisées est plus ou moins discutable puisque la stylisation de l'émotion n'est pas toujours comparable à l'expression de l'émotion qui est vécue par le locuteur ou la locutrice dans une situation réelle. L'émotion de détresse, analysée içi, montre les aspects de stress et de tristesse. Le premier se manifeste par l'augmentation du F0 moyen, et le dernier par le ralentissement du débit (voir [Str83] et [Sch86).

L'expérience I montre que l'émotion est mieux perçue dans la partie finale que les autres parties, initiale et médiane, de la phrase. Le F0 moyen est un bon indice prosodique pour la perception de l'émotion de détresse. Notre résultat est en accord avec la conclusion de Scherer et al. [Sch84] sur la nécessité de l'intergration des deux modèles dans le recherche de la parole. C'est-à-dire, la perception différente de l'émotion selon parties différentes de la phrase montre l'interaction des niveaux linguistique et paralinguistique qui est proposée par le modèle de la configuration alors que corrélation entre la valeur de F0 moyen et le taux de la perception de l'émotion soutient le modèle de la covariation.

L'expérience II met en lumière l'unité minimale de la perception de l'émotion dans le domaine temporel. La durée de 100ms, la durée moyenne d'une voyelle dans cette expérience, semble être minimale pour que l'auditeur puisse percevoir l'émotion du locuteur correctement. Cette durée parait suffisante pour la perception de l'émotion dans cette expérience. Cependant, cette expérience n'indique pas si cette durée minimale pour la perception de l'émotion se réfère à la durée absolue de 100ms ou à la durée d'une voyelle. D'autres expériences sont nécessaires pour répondre à cette question. L'indice prosodique corrélé avec le taux de la perception de l'émotion est le F0 moyen pour les stimuli de phrases ou de parties de la phrase, alors que c'est la perturbation de F0 pour les stimuli de la durée minimale. Cette différence suggère une stratégie différente de la perception de l'émotion selon le type de stimuli ou la durée de stimuli.

Le taux de perception relativement élevé de l'émotion à travers les locutrices et même dans les stimuli minimaux indique la stabilité de la perception de l'émotion malgré la variation dans les stratégies. Cette stabilité contraste avec l'instabilité des indices acoustiques de l'émotion en général. Streeter et al. [STR83] notent que le comportement de l'auditeur dans la perception de l'émotion est plus prédictible que la manifestation acoustique de l'émotion.

Remerciement

L'auteur tient à remercier Philip Lieberman, professeur au département des sciences cognitives et de linguistique de l'université de Brown, pour des discussions fructueuses au cours des expériences et John Mertus du même département pour son support technique. Elle remercie également Jacqueline Vaissière, professeur au département de phonétique à l'université de Paris III, de ses commentaires précieux sur le travail.

Bibliographie

[Fon83] Fonagy, I. (1983), *La vive voix*, Payot, Paris.

[Kai62] Kaiser, L. (1962), "Communication of affect by single vowels," *Synthese* 14, 300-319.

[Léo93] Léon, P. R. (1993), *Précis de Phonostylistique : Parole et expressivité*, Nathan, Paris.

[Lei97] Leinonen, L. *et al.* (1997), "Expression of emotional-motivational connotations with a one-wordutterance,"*J.A.S.A.*102(3),1853-63.

[Pro97] Protopapas, A. and Lieberman, P. (1997). Fundamental frequency of phonation and perceived E stress. *J. Acoust. Soc. Am.*

[Sch72] Scherer, K. R. *et al.* (1972), "Minimal cues in the vocal communication of affect," *J. of Psycholinguistic Research* 1, 269-285.

[Sch77] Scherer, K. R. and Oshinsky (1977), "Cue Utilization in Emotion Attribution from Auditory Stimuli," *Motiv.&Emot.*1(4),331-46.

[Sch84] Scherer, K. R. *et al.* (1984), "Vocal cues to speaker affect", *J. A. S. A.* 76(5), 1346-1356.

[Sch86] Scherer, K. R. (1986), "Vocal Affect Expression:," *Psychol. Bull.* 99 (2), 143-165.

[Str83] Streeter, L. A., *et al.* (1983), "Acoustic and Perceptual Indicators of Emotional Stress," *J. Acoust. Soc. Am.* 73(4), 1354-1360.

[Vai83] Vaissière, J. (1983), "Language-Independent Prosodic Features", in the *Prosody: Models and Measurements*, by Sprigner-Verlag,53-66.

Pertinence perceptive des configurations intonatives en parole émotionnelle

Sylvie J. L. Mozziconacci et Dik J. Hermes

IPO
P.O.Box 513, 5600 MB Eindhoven, The Netherlands
E-mail: mozzic@ipo.tue.nl - Tél.: +31 40 2475200

ABSTRACT

In a previous production study, the pitch contours of emotional utterances were labelled in terms of the IPO intonation grammar and hypotheses were drawn concerning the role of the patterns in emotional speech. In the present perception study, these hypotheses were tested. The listening test provided converging evidence on the contribution of specific intonation patterns in the perception of some of the emotions and attitudes studied. Some intonation patterns such as final '3C' and '12', that were specifically produced in some emotion, also introduced a perceptual bias towards that emotion. In that sense, the results from the perception study supported the results from the previous production study.

1. INTRODUCTION

L'importance qu'ont les paramètres prosodiques tels que la hauteur mélodique (*pitch level*), la dynamique fréquentielle (*pitch range*) et le débit de parole pour, entre autres, l'expression de l'émotion dans la parole, est reconnue dans diverses études [Wil72, Sch89, Cah90, Car92, Kit92, Vro93]. Les études consacrées à l'expression de l'émotion et de l'attitude dans la parole [Fri85, Mur93] se sont rarement penchées sur la relation entre la production et la perception, et se sont généralement limitées à l'étude de paramètres abordés sur une base de statistiques. La présente étude est consacrée au rôle des configurations intonatives (*intonation patterns*) dans la perception de l'émotion et de l'attitude dans la parole. Dans le présent texte, le terme "émotion" est utilisé pour couvrir sans distinction les deux notions d'émotion et d'attitude.

2. CONTEXTE ET BUT DE CETTE ÉTUDE

Dans l'étude précédente, sept émotions : la neutralité (catégorie de référence), la joie, l'ennui, la colère, la tristesse, la peur et l'indignation, ont été simulées par trois locuteurs hollandais, deux hommes et une femme. Pour se mettre en situation, les locuteurs ont d'abord dit un texte de contenu sémantique émotionnellement chargé correspondant à l'émotion à exprimer. Imprégnés de cette situation émotionnelle, ils ont ensuite dit cinq textes considérés de contenu sémantique neutre en exprimant cette même émotion. La procédure a été répétée 3 fois, pour l'enregistrement d'un total de 315 phrases (3 fois * 5 textes * 7 émotions * 3 locuteurs). Ces phrases ont ensuite été étiquetées par deux phonéticiens expérimentés en étiquetage selon la grammaire intonative de l'IPO [Har90]. Leurs réponses ont été distribuées dans des catégories de configurations intonatives grammaticales. Les configurations intonatives de cette grammaire sont composées de dix mouvements mélodiques de base. En dehors du fait qu'ils sont montants ou descendants, ces mouvements sont également caractérisés à l'aide de leur *timing* dans la syllabe, de leur taille, et de leur extension sur une ou plusieurs syllabes. Il s'agit des mouvements de base suivants : '1' une montée mélodique précoce et conférant un accent, '2' une montée mélodique très tardive ne conférant pas d'accent, '3' une montée mélodique tardive conférant un accent, '4' une montée progressive étendue sur plusieurs syllabes, '5' une semi-montée mélodique excédant la montée mélodique qui la précède (*overshoot*) et conférant un accent, 'A' une descente mélodique tardive conférant un accent, 'B' une descente mélodique précoce ne conférant pas d'accent, 'C' une descente mélodique très tardive ne conférant pas d'accent, 'D' une descente mélodique lente étendue sur plusieurs syllabes, 'E' une semi-descente mélodique pouvant conférer un accent. Deux mouvements réalisés sur une même syllabe sont liés par le symbole "&" dans l'étiquetage, comme par exemple dans "1&A". Toutes les combinaisons de ces mouvements ne sont pas considérées comme étant grammaticales. Pour le néerlandais, les principales catégories grammaticales sont les suivantes : 1&A (montée puis descente sur une seule syllabe), appelé '*pointed hat*' par analogie de forme, 1A (montée sur une syllabe accentuée et descente sur la syllabe accentuée suivante) aussi nommé '*flat hat*', 1B (montée sur une syllabe accentuée et descente précoce, précédant généralement une frontière phonologique), 1D (montée sur une syllabe accentuée et descente progressive), 5A (*overshoot* suivant une montée, et descente), 1EA (montée, semi-descente, et descente), 12 (montée suivie d'une montée très tardive), 1C (montée et descente très tardive), 3C (montée tardive et descente très tardive), 45&A (montée progressive, *overshoot* et descente sur une syllabe unique).

Les résultats suggèrent que le choix de configuration intonative est d'une pertinence toute particulière pour l'expression de certaines émotions dans la parole. Même si certaines configurations intonatives transmettent l'émotion mieux que d'autres, une configuration intonative particulière ne paraît pas absolument nécessaire

pour signaler l'expression de l'émotion dans la parole ; il n'y a pas non plus de correspondance unique entre une émotion et une configuration intonative. On constate aussi que, même si une configuration intonative autre que '1&A' a été réalisée le plus fréquemment lors de l'expression de certaines émotions, '1&A' a été employé dans l'expression de chacune des émotions. Une première hypothèse consiste à considérer '1&A' comme adéquat pour exprimer toutes les émotions considérées ici. De plus, dans la neutralité, les locuteurs n'ont réalisé aucun '12', '1&2' et 'C' en position finale. Une deuxième hypothèse consiste à dire que ces configurations intonatives peuvent servir à signaler l'émotion dans la parole.

Une expérience a été faite afin de vérifier si certaines configurations intonatives contribuent à la perception de certaines émotions dans la parole. Des contours mélodiques ont été synthétisés en employant une série de configurations intonatives différentes. Ces contours ont été projetés sur des phrases neutres. Les sujets devaient tenter d'identifier l'émotion perçue à l'écoute des phrases ainsi obtenues. Si l'emploi de différentes configurations intonatives entraîne une différence significative dans la distribution des réponses des sujets, la pertinence du choix de configuration intonative sera établie pour l'expression de l'émotion dans la parole.

3. MATÉRIEL

Deux phrases neutres, l'une sur le Texte 1 : "Ze hebben een nieuwe auto gekocht" (Ils ont acheté une nouvelle voiture), l'autre sur le Texte 2 : "Zijn vriendin kwam met het vliegtuig" (Son amie est venue en avion) ont été manipulées par analyse-resynthèse. Les manipulations ont combiné chacune des configurations intonatives participant à cette étude, avec un *pitch level* et un *pitch range* appropriés à l'expression d'une émotion particulière. Les valeurs de *pitch level* et de *pitch range* utilisées pour une émotion donnée (Table 1) ont été trouvées optimales à l'issue d'une série d'expériences faites dans une étude précédente [Moz94]. Quant aux configurations intonatives, seules celles qui sont considérées grammaticalement correctes selon la grammaire intonative de l'IPO [Har90] et qui ont été réalisées plus qu'occasionnellement dans les enregistrements étudiés précédemment font l'objet de la présente étude. Il s'agit des 11 configurations intonatives suivantes : 1&A 1&A, 1B 1&A, 1D 1&A, 12 1&A, 15&A, 1A, 1EA, 1&A 3C, 1B 3C, 1D 3C, and 1&A 12. Des points d'ancrage ont été choisis pour la synthèse des mouvements mélodiques. D'autres études suggèrent que le *vowel onset* est plus approprié que le début du voisement en tant qu'instant d'initialisation du mouvement mélodique dans la syllabe [Ale95]. Ceci est lié au fait que les parties stables du spectre jouent un rôle proéminent dans l'intonation [Hou90], et que la partie consonantique précédant la voyelle est plus sujette à de rapides variations spectrales que ne l'est la voyelle. D'autre part, les parties voisées suivant la voyelle dans la syllabe peuvent être pertinentes pour effectuer certaines distinctions linguistiques [Her97]. Par conséquent, les points d'ancrage utilisés sont les *vowel onsets* des syllabes portant l'accent lexical, et la fin du voisement. Le timing et la durée des mouvements mélodiques pour la synthèse des '1&A' et des 'A' ont été dérivés des descriptions de Collier [Col91]. La réalisation des autres mouvements mélodiques a été basée en partie sur la description de mouvements mélodiques représentatifs donnée par 't Hart et Collier [Har75]. Pour certains mouvements mélodiques, cette description étant incomplète, les contours mélodiques réalisés par les locuteurs ont servi de point de départ pour trouver des spécifications adéquates. Ces spécifications seront présentées par Mozziconacci [MozFo].

Un total de 154 stimuli a ainsi été créé (11 configurations intonatives * 2 textes * 7 combinaisons de *pitch level* et de *pitch range*). Les manipulations sont basées sur la technique PSOLA [Mou95] et n'ont pas affecté la durée segmentale.

Table 1 : Valeurs de *Pitch level* et de *pitch range* trouvées optimales dans l'étude précédente.

	Pitch level (Hertz)	*Pitch range(semitones)*
Neutralité	65	5
Joie	155	10
Ennui	65	4
Colère	110	10
Tristesse	102	7
Peur	200	8
Indignation	170	10

4. PROCÉDURE

Les stimuli ont été présentés en deux différents blocs ; un bloc par texte. La moitié des sujets a commencé par le texte 1, l'autre par le texte 2. A l'intérieur de ces blocs, les stimuli ont été présentés à chaque sujet dans un ordre aléatoire différent. Au total, 24 sujets (12 hommes et 12 femmes) ont participé à ce test. Aucun d'entre eux n'avait de connaissances particulières en phonétique. Ils ont participé au test séparément, en utilisant un casque audio et un programme interactif. Les sujets ne pouvaient écouter chaque phrase qu'une seule fois et n'ont reçu aucun feed-back concernant leurs performances. Ils ont eu une pause entre les deux blocs, séparant le test en deux périodes, chacune d'environ 10 à 15 minutes. Leur tâche consistait à attribuer à chaque stimulus l'une des sept émotions données, en l'occurrence l'émotion qui, selon eux, avait été exprimée dans la phrase écoutée.

5. RÉSULTATS

Pour chaque configuration intonative, les réponses des sujets ont été additionnées par catégorie de réponses. Les distributions de réponses obtenues pour chaque configuration intonative ont été comparées entre elles ; pour ce faire 110 χ^2 ont été calculés. Afin d'apprécier les

différences de perception de l'émotion en fonction de la configuration intonative employée, les deux χ^2 correspondant à la même paire de configurations intonatives ont été moyennés, puis classés par ordre croissant. Sur cette base, des groupements distincts de configurations intonatives peuvent être constitués dans un espace à deux dimensions. Les configurations intonatives formant de tels groupements sont représentées séparément pour le Texte 1 (Figure 1), et le Texte 2 (Figure 2). L'ordre des distances y est spécifié grâce à un chiffre. Plus le chiffre est élevé, plus la distance entre les configurations est importante. Les différences non-mentionnées dans ces figures correspondraient à d'importantes distances puisque les différences entre configurations intonatives étaient très significatives ($p < 10^{-4}$). La configuration intonative '1&A 12' était clairement différente des autres, puisque pour tous les χ^2, $p < 10^{-9}$.

Un groupement est constitué de : '1&A 1&A', '1B 1&A', '1D 1&A' et '12 1&A'. Les configurations : '1&A 3C', '1B 3C' et '1D 3C' forment un autre groupe. '1A' et '1EA' en constituent un troisième. Par conséquent, c'est surtout la partie finale de la configuration intonative qui affecte la perception des émotions dans la parole.

Il s'avère que, pour l'ensemble des configurations intonatives, l'identification de l'ennui et de la neutralité était très élevée, alors que l'identification de la tristesse et de la colère était faible. Il s'avère aussi que pour une combinaison de *pitch level* et de *pitch range* donnée, combinaison considérée optimale pour une émotion particulière, l'emploi d'une configuration intonative spécifique influence l'identification de l'émotion. Quand '1&A 1&A', '1B 1&A', '1D 1&A' et '12 1&A' sont employés, les sujets ont tendance à répondre que l'émotion exprimée est la neutralité. L'utilisation de '1&A 12' a occasionné la plus faible identification de l'ennui. L'emploi de '1A' et de '1EA' semble augmenter le nombre de réponses dans la catégorie indignation, et l'emploi de '1EA' diminuer le nombre de réponses dans la catégorie peur. L'emploi de '1A' et de '1EA' fournit les résultats les meilleurs pour l'expression de la colère. Bien que l'identification de la tristesse soit faible et l'identification de la peur médiocre, les résultats les meilleurs sont obtenus en utilisant '3C' en position finale. De plus, l'emploi de cette configuration intonative induit chez les sujets une tendance à répondre que l'émotion exprimée est l'indignation quand le pitch est élevé et à opter pour l'ennui quand le pitch est bas. L'utilisation de '1&A 12' influence les sujets en faveur de l'indignation. Quand '12' et, dans une certaine mesure, quand '3C' ont été utilisés en position finale, même les phrases portant les valeurs optimales de pitch pour l'expression de la neutralité ont été perçues comme émotionnelles. Ces configurations de mouvements mélodiques ne sont donc pas adéquates pour exprimer la neutralité.

Les résultats montrent que le choix de la configuration intonative est pertinent à la perception de l'émotion dans la parole. Les sujets ont tendance à percevoir certaines émotions plutôt que d'autres, en fonction de la configuration intonative employée.

6. Discussion générale

L'expérience de perception a démontré qu'une synthèse employant '1&A 1&A' comme unique configuration intonative combinée avec des variations adéquates en *pitch level* et *pitch range* donne des résultats acceptables pour toutes les émotions étudiées. Il s'avère que, pour les cas dans lesquels il est souhaitable que les configurations intonatives ne contribuent pas à la variabilité observée lors d'une expérience, '1&A' permet de contrôler cette variabilité en configurations intonatives. Les résultats montrent que le choix de configuration intonative est tout à fait pertinent pour la perception de l'émotion dans la

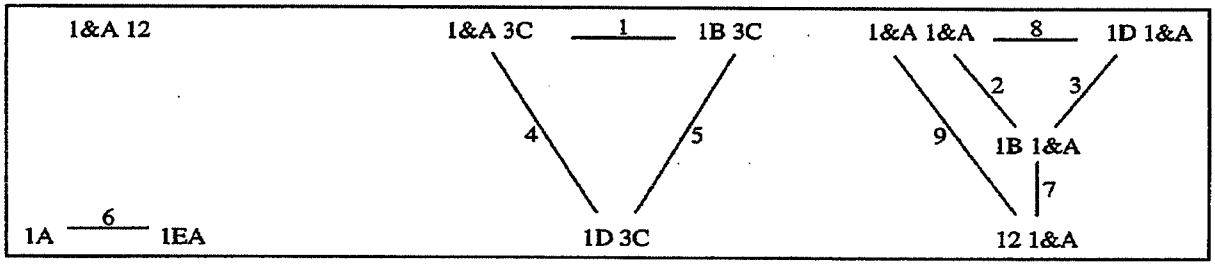


Figure 1 : Texte 1 : Ze hebben een nieuwe auto gekocht. (Ils ont acheté une nouvelle voiture.)

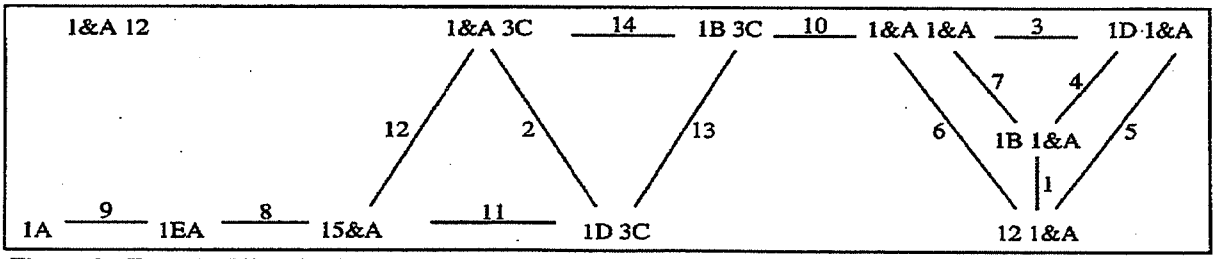


Figure 2 : Texte 2 : Zijn vriendin kwam met het vliegtuig. (Son amie est venue en avion.)

parole. Il s'avère qu'une configuration intonative spécifique est plus adéquate pour exprimer certaines émotions que pour en exprimer d'autres. En particulier, '12' en position finale influence fortement la perception des sujets en faveur de l'indignation. Les deux hypothèses dérivées de l'étude de la production ont donc été vérifiées. Enfin, les résultats concernant la production et ceux concernant la perception correspondent bien puisqu'il se trouve que les configurations intonatives réalisées le plus fréquemment par les locuteurs dans l'expression de certaines émotions, influencent la perception des sujets en faveur de ces mêmes émotions.

Une analyse détaillée, qui sera développée dans [MozFo] révèle les correspondances suivantes entre émotions et configurations intonatives en position finale dans la phrase. Pour exprimer la neutralité, 1&A est la configuration à préférer. Si une autre configuration est employée, ce ne doit être ni 3C ni 12. L'ennui est réalisé en employant de préférence 3C, et en évitant l'emploi de 5&A ou de 12. La configuration 3C favorise la perception de la tristesse, alors que 5&A y est défavorable. Pour la peur, 12 est la configuration préférable, alors que A ou EA sont à éviter. Pour la joie, l'emploi de 1&A ou de 5&A est indiqué, alors que l'emploi de A, EA ou 12 est à déconseiller. Pour la colère, les configurations 5&A, A ou EA sont indiquées et les configurations 1&A et 3C deconseillées. La configuration 12 est fortement associée à l'expression de l'indignation. C'est donc la configuration à utiliser de préférence pour cette émotion, alors que 1&A est à éviter.

Enfin, puisque les configurations intonatives s'avèrent pertinentes pour l'expression de l'émotion dans la parole, les traits distinctifs de ces mouvements mélodiques tels qu'ils sont définis dans la grammaire intonative de l'IPO [Har90], doivent être pertinents pour la communication.

Remerciements

Ces recherches ont été effectuées dans le cadre d'un programme de collaboration entre les universités du Brabant (SOBU). Nous souhaitons exprimer notre gratitude envers Don Bouwhuis pour ses précieux conseils concernant l'analyse des résultats.

Bibliographie

[Ale95] d'Alessandro C., and Mertens P. (1995), "Automatic pitch contour stylization using a model of tonal perception", Computer Speech and Language, Vol. 9, pp. 257-288.

[Cah90] Cahn J.E. (1990), Generating expression in synthesized speech, MIT Media Lab. Techn. Report.

[Car92] Carlson R., Granström B., and Nord L. (1992), «Experiments with emotive speech: acted utterances and synthesized replicas», ICSLP 92, pp. 671-674.

[Col91] Collier R. (1991), "Multi-language intonation synthesis", Journal of Phonetics, Vol. 19, pp. 61-73.

[Fri85] Frick R.W. (1985), "Communicating emotion: the role of prosodic features", Psychological Bulletin, Vol. 97, pp. 412-429.

[Har75] Hart J. 't, and Collier R. (1975), "Integrating different levels of intonation analysis", Journal of Phonetics, Vol. 3, pp. 235-255.

[Har90] Hart J. 't, Collier R., and Cohen A. (1990), A perceptual study of intonation, Cambridge University Press.

[Her97] Hermes D.J., Beaugendre F, and House D. (1997), "Individual differences in accentuation boundaries in Dutch", IPO Annual Progress Report, Vol. 32, pp. 131-138.

[Hou90] House D. (1990), Tonal perception in speech, Lund University Press.

[Kit92] Kitahara Y., and Tohkura Y. (1992), "Prosodic control to express emotions for man-machine interaction", IEICE Trans. Fundamentals, Vol. 75, pp. 155-163.

[Moz94] Mozziconacci S.J.L. (1994), Pitch and duration variations conveying emotions in speech, IPO Report No. 961.

[MozFo] Mozziconacci S.J.L. (forthc.), "Some variations due to the expression of emotion and attitude in speech", in preparation.

[Mou95] Moulines E., and Laroche J. (1995), "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Communication, Vol. 16, pp. 175-205.

[Mur93] Murray I.R., and Arnott J.L. (1993), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", J. Acoust. Soc. Am., Vol. 93-1, pp. 1097-1108.

[Sch89] Scherer K.R. (1989), "Vocal measurement of emotion", In R. Plutchik and H. Kellerman (Eds.), Emotion: theory, research, and experience, Vol. 4, Academic Press, San Diego.

[Vro93] Vroomen J., Collier R., & Mozziconacci S. (1993), «Duration and intonation in emotional speech», Eurospeech' 93, pp. 577-580

[Wil72] Williams C.E., and Stevens K.N. (1972), "Emotions and speech: some acoustical factors", J. Acoust. Soc. Am., 52, pp. 1238-1250.

La variabilité du rythme cardiaque et la différentiation prosodique des émotions

Branka Zei et Marc Archinard

Hôpital Universitaire de Genève, Psychiatrie de Liaison
61 Bd. De la Cluse 1205 Genève
Tél.: ++41 22 382 48 81- Fax: ++41 27 721 48 99
e-mail: branka-zei@diogenes.hcuge.ch

ABSTRACT

In order to investigate the neurological substratum of prosodic marking of emotional expression, heart rate variability - as indicator of autonomous nervous system reactivity - was measured and correlated with a composite index of vocal arousal reflecting vocal differentiation of emotionally marked utterances.

The results show that subjects with low heart rate variability do not exhibit prosodic differentiation between vocally expressed anger/joy and sadness.

1. INTRODUCTION

Les émotions ont été décrites comme des processus psychophysiologiques qui englobent des cognitions, des changements humoraux et immunologiques, des expressions verbales et non verbales aussi bien que des dispositions comportementales telles que: approche, évitement répulsion. Ces dernières sont soutenue par l'action du système nerveux autonome (SNA).

La relation entre le SNA et la parole a été modélisée par Williams C.E. et Stevens N. [Wil81], comme suit (paraphrasé par l'auteur): "L'activation accrue de la branche sympathique du SNA apparaît lors de l'expérience de colère, peur ou joie intense. Les conséquences en sont les suivantes: une augmentation du rythme cardiaque, de la tension artérielle et la distribution du sang dans la musculature striée.

Ces changements physiologiques entraînent des modifications de la profondeur et de la fréquence respiratoire, aussi bien qu'une diminution des sécrétions salivaires et le dessèchement de la bouche. L'activation accrue du système parasympathique conduit à la diminution de la fréquence cardiaque, à la réduction de la pression sanguine et l'éloignement du sang de la musculature striée lors de l'expérience du chagrin. Ces effets physiologiques influencent le fonctionnement des appareils phonatoires, articulatoires et respiratoires, et provoquent des changements de position de sternum.

L'ensemble de ces changements physiologiques avec des dispositions comportementales y relatives, sont connus sous le terme de "emotional arousal"- excitation émotionnelle. Nous avons baptisé "vocal arousal"- excitation vocale, l'ensemble des caractéristiques de la parole relatives à l'excitation émotionnelle.

L'équilibre entre les actions des deux systèmes du SNA (sympathique et parasympathique) a été abondamment étudié au travers la variabilité de la fréquence cardiaque (VFC). Cette dernière est considérée comme indice fiable de la réactivité de base du SNA. Un niveau plus élevé de la VFC est associé au fonctionnement émotionnel normal, alors qu'une faible VFC est associée à des troubles affectifs tels que attaques de panique ou névrosisme. La parole étant influencée par la psychophysiologie, nous avons examiné son lien avec la VFC en énonçant les hypothèses suivantes:

1. Le réactivité de base du SNA (mesurée par la variabilité de la fréquence cardiaque (VFC) influence le degré d'excitation émotionnelle de sorte qu'une faible VFC provoque un aplatissement des excitations émotionnelles.

2. L'aplatissement des excitations émotionnelles provoque un aplatissement des réactions vocales (vocal arousal).

3. La réaction vocale étant en général plus forte dans la colère que dans la tristesse [Ban96], les sujets à faible VFC ne motteraient pas cette différence.

2. PLAN EXPERIMENTAL

2.1 Sujets

Pour obtenir des sujets susceptible d'avoir une diminution la VFC, nous avons examiné 40 patients diabétiques d'âge et de durée de la maladie variables. Le choix des sujets a été motivé par le fait que le diabète provoque souvent des lésions du SNA (neuropathie autonome).

2.2 Données physiologiques

Deux tests standard de la VFC [Vit86], ont été effectués: 1. Le test de respiration profonde et 2. Le test de changement de posture.

2.3 Induction de l'excitation émotionnelle et des réactions vocales y relatives

Selon les travaux de Mendolia [Men93], l'évocation verbale des situations émotionnelles dirige l'attention vers les contenus à charge affective. Le récit même stimule les images mentales et provoque le sentiment émotionnel au moment du récit [Tes78]. Aussi une expérience émotionnelle passée est rendue présente par la parole [Str85].

2.4 Données psychologiques

Les échelles d'anxiété-état [Bru93], et de personnalité [Eys71], ont été utilisée pour évaluer d'éventuelles interférences des ces facteurs psychologiques.

2.5 Procédure

On a demandé à 40 patients diabétiques de raconter des situations émotionnelles réellement vécues qui leur ont fait ressentir de la colère, de la joie et de la tristesse. Au terme de chaque récit le sujet a prononcé la phrase standard: "ALORS TU ACCEPTES CETTE AFFAIRE".

La phrase a été présentée par écrit, en lettres majuscules et sans ponctuation pour éviter une suggestion quant à l'interprétation prosodique de la phrase. Les voix des patients ont été enregistrées sur une bande DAT. La distance (4cm de la bouche) et l'orientation du microphone ont été tenus constants moyennant un support placé sur la tête.

3. ANALYSES ACOUSTIQUES

3.1 Méthodes de mesure des paramètres vocaux relatifs à l'excitation émotionnelle

Toute les analyses ont été faites avec "Signlyze", un logiciel fonctionnant sur Macintosh [Kel95]. Trois catégories d'indices vocaux ont été extraites:

1. La fréquence fondamentale calculée à partir des signaux digitalisés à 16KHz
2. L'énergie acoustique exprimée en pseudo-décibels.
3. La vitesse d'élocution exprimée en nombre de syllabes prononcées dans une seconde.

Calcul des paramètres de la F0 La routine FFT-Peigne a été utilisée pour l'extraction de la fréquence fondamentale. Elle "exécute des FFT successives en 512 points, et les informations fréquentielles provenant à la fois de la fréquence fondamentale et des harmoniques sont combinées pour déterminer le contour de la F0" [Kel95p.185] .L'intervalle du calcul du spectre à été fixée à 5 ms. Un filtre de sortie de 8 a été utilisé pour éliminer des valeurs "extrêmes". Le seuil de détection de la présence/absence du voisement a été adapté aux caractéristiques de chaque signal pour optimiser le détection des portions voisées.

Les paramètres suivants ont été calculés à partir des courbes de la fréquence fondamentale: les valeurs moyenne et médiane, l'écart 5^{e} -95 ^{e} centile, le rapport Max/min, l'écart type de la moyenne, le coefficient de variation.

Calcul des paramètres de l'énergie La routine RMS (racine carrée de la moyenne) dessine l'enveloppe sur la base d'une moyenne "glissante" de la fenêtre. La largeur de la fenêtre a été fixée à 50 ms. Les paramètres suivants ont été extraits des enveloppes de l'énergie: énergie voisée maximum, la moyenne de énergie voisée, l'écart entre les pics maximum et minimum de l'énergie voisée. Les points de mesure étaient relevés aux centres des noyaux vocaliques. Figure 1 illustre les points de mesure de l'écart dB_pic-à-pic.

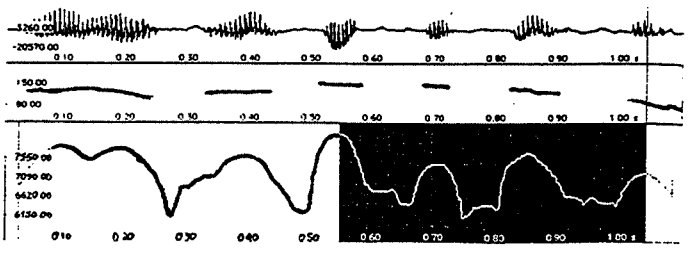


Figure 1: Les points de mesure de l'écart entre énergie voisée maximum et minimum.

Calcul de la vitesse d'élocution L'expérimentateur a compté "à l'oreille" le nombre de syllabes prononcées par le sujet. Ce nombre a varié selon le style vocal et/ou l'accent de chaque sujet. Comme la phrase standard comportait deux groupes syntagmatiques (chacun ayant une forte cohésion linguistique et constituant potentiellement une courbe intonative complète), les sujets introduisaient des pauses de durée 0.1 - 2 sec. entre les deux unités syntagmatiques. De telles pauses ont été coupées et la durée totale du signale a été mesurée ensuite. La vitesse d'élocution a été calculée comme suit: le nombre de syllabes/la durée du signal.

Aussi, la vitesse d'élocution correspondait plutôt à la vitesse de l'articulation qu'à la vitesse générale d'élocution. Cette façon de calculer convenait à la mesure l'accélération et de décélération articulatoire lors de l'expression de la colère ou de la tristesse.

3.2 Transformations des données vocales

Dans le but d'éliminer les différences vocales dues aux sexe toutes les valeurs ont étés normalisées par rapport aux valeurs de référence externes obtenues à partir des analyses des voix d'acteurs prononçant la même phrase type sur un ton neutre [Sig95]. L'accord inter-juges obtenu pour cette phrase a été au dessus de 95%. Les valeurs de référence ont été obtenues dans notre laboratoire d'analyse de la voix en appliquant les mêmes procédures d'analyses que celles utilisée pour les voix des patients diabétiques.

3.3 Transformations des données physiologiques

Les scores obtenus sur les tests de la VFC ont été normalisés par référence aux valeurs normales attendues pour chaque âge. Le cumul des scores sur les deux tests a constitué l'Index_VFC pour chaque patient.

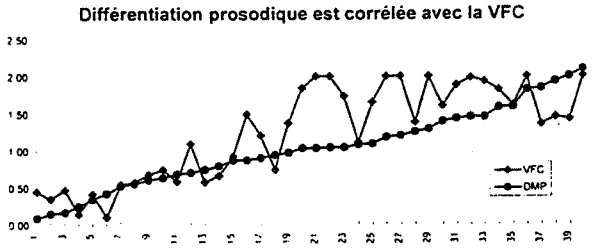
4. ANALYSES STATISTIQUES ET RESULTATS

4.1 Création de l'Indice de saillance prosodique

Les analyses statistiques (régressions multiples méthode pas-à-pas) ont mis en évidence trois paramètres vocaux pertinents par rapport à la VFC: F0Max/min, l'écart dB_pic-à-pic, et la vitesse d'élocution. Nous avons cumulé les trois paramètres en constituant un indice global de saillance prosodique (ISP).

4.2 Création de l'Indice de différentiation prosodique maximale

Pour saisir la capacité de nos patients de différencier vocalement les émotions fortement opposées quant au degré d'activation du SNA, nous avons calculé le delta entre l'ISP_colère et l'ISP_tristesse. Le même calcul a été fait pour la différence entre le ISP_joie - ISP_tristesse. La valeur la plus grande des deux a été retenue pour constituer l'indice de différentiation prosodique maximale (DPM). Les corrélations partielles entre l'indice de la VFC et le DPM (contrôlées pour l'anxiété et l'extraversion) ont été statistiquement significatives: r = .76; P=.000. La figure 2 illustre la corrélation positive entre la différentiation prosodique des émotions et l'indice de la variabilité de la fréquence cardiaque.



Des analyses des régressions multiples (méthode pas-à-pas , alpha = .05) avec le DPM comme variable dépendante et les variables psychologiques et physiologiques comme indépendantes ont donné les résultats suivants:

La VFC seule explique 60% de la variance avec R^2 = 60; R Multiple = 77; T = 7.6; P = .0000.

Ces résultats nous permettent de conclure que l'expression vocale des émotions et le degré de la différentiation prosodique des états affectifs (étudiés dans notre expérience c.-à-d. joie, colère, tristesse) sont liés à la réponse de base du SNA. Plus le système réagit de façon dynamique (reflétée dans la variabilité de la fréquence cardiaque) plus les sujets parviennent à différencier vocalement leurs états affectifs. Aussi l'aplatissement de l'excitation émotionnelle induit-il un aplatissement du marquage prosodique des énoncés.

BIBLIOGRAPHIE

[Ban96] *Banse R., & Scherer K.R. 1996; "Acoustic Profiles in Vocal Emotion Expression", Journal of Personality and Social Psychology, Vol. 70, No 3, 614 - 636*

[Bru93] *Bruchon-Schweitzer, M. (1993). C.D. Spielberger: Inventaire d'anxiété état-trait, Forme Y. Editions du Centre de Psychologie Appliquée*

[Eys70] *Eysenck, H.J. (1970),. The structure of human personality,. London: Methuen.*

[Kel95] *Keller, E. (1994). Signal Analysis for Speech and Sound. Lausanne: Info Signal, Inc.*

[Men93] *Mendolia M., Kleck R.E. 1993; "Effects of Talking About a Stressful Event on Arousal: Does What We Talk About Make a Difference?" Journal of Personality and Social Psychology, Vol 64, No 2, 283 - 292*

[Sig95] *Siegwart, E., 1993, Traitement des fonctions expressive et linguistique de la prosodie: une étude de neuropsychologie, Thèse présentée à la Faculté de Psychologie er des sciences de l'education de l',Université de Genève; thèse No 199*

[Str78] *Strack, F., Schwarz, N., & Gschneidinger, E. 1985; "Happiness and reminescing: The role of time perspective, affect, and mode of thinking" ;Journal of Personality and Social Psychology, 49, 1460 - 1469*

[Tes78] *Tesser, A., Leone C., & Clary, G. 1978; "Affect control: Process contraints versus catharsis", Cognitive Therapy and Research, 2, 265 - 274*

[Vit86] *Vita G., Princi, P., (1986); "Cardiovascular Reflex tests", Journal of the Neurological Sciences, 1986,75: 263-274, Elsevier Sycience Publishers B.V.*

[Wil81] *Williams, C. E., & Stevens, K.N. (1981). "Vocal Correlates of Emotional States", Speech Evaluation in Psychiatry. 221-240. New York: Darby, J.K. Grune, & Statton.*

Ressources

Spécification, réalisation et validation d'un corpus oral pour la reconnaissance de la parole dans une voiture

Christelle Gassert et Jean-François Mari

LORIA & INRIA-Lorraine
Campus scientifique
B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX

RÉSUMÉ

Many accessories manufacturers are studying the integration of a speech recognition system in a car. Speech recognition in a car can enhance the security around the driving task because the driver can keep his attention focused on the road when he is activating various accessories by voice. The study of the interaction, the specification and the training of the speech recognition system involve the recording *in situ* of a corpus of various requests in the presence of noise. Speech recognition in a car encounters several issues. One of the most important is the presence of noise coming from various origins: motor, pavement, other accessories and crosstalk. In order to cope with this difficulty and to train a speech recognition system, we have recorded a vast speech database in a car and carried out a classification of the sources of noise. This classification is used to annotate the speech signal in addition to the labeling into words. In this paper, we present our work in specifying and recording a new vast speech database designed to train a speech recognition system capable to activate various accessories like: a telephone, a radio, and a navigation system. A detailed attention is devoted toward the automatic labeling of the speech signal.

1. INTRODUCTION

Les progrès du traitement de la parole posent aujourd'hui la question de l'introduction de ces technologies dans les véhicules automobiles de demain. La reconnaissance de la parole s'est longtemps heurtée à des problèmes techniques et économiques. Avec les progrès de la recherche en reconnaissance automatique de la parole, de l'électronique et la baisse des coûts des composants, nous pouvons aujourd'hui imaginer que les prestations classiques telles l'autoradio ou le téléphone pourraient être commandées par la parole, ainsi que les nouvelles prestations comme la navigation.

Commander des accessoires par la voix est présenté comme une facilité d'utilisation qui augmente la sécurité de la conduite. Le conducteur peut garder ses mains disponibles pour la conduite ; il peut aussi, le cas échéant, allier le geste à la parole en utilisant un clavier simplifié permettant l'accès à des menus déroulants.

Le projet européen VODIS (Voice Operated Driver Information System) de la branche Telematics s'intéresse à l'intégration d'un système complexe de navigation par satellites à l'intérieur d'un véhicule et à sa commande par la voix. Dans cette application, le conducteur peut demander oralement la sélection d'une source de musique, d'un numéro de téléphone ou programmer le système de navigation GPS (Global Positional System) qui l'assiste dans le choix de la route.

La réalisation d'un tel système pose des problèmes scientifiques qui sont encore du domaine de la recherche. Citons :

- la séparation de sources (*speech separation*) [1] afin d'isoler la parole du conducteur des autres conversations ;
- la valorisation de la parole du conducteur (*speech enhancement*) afin de supprimer le signal issu de l'autoradio ;
- la reconnaissance dans le bruit particulièrement nonstationnaire[8, 7] ;
- l'intégration du reconnaisseur dans un système multimodal.

La résolution de ces problèmes passe par l'enregistrement d'un corpus de parole bruitée à l'aide de plusieurs microphones répartis dans le véhicule.

Ce projet, dont la direction est assurée par la société Bosch qui a construit le système de navigation Berlin RCM 303A, regroupe les principaux constructeurs automobiles européens. L'INRIA - Lorraine est le partenaire associé de Renault-Innovation chargé de la réalisation du corpus de commandes orales exprimées en français et destiné à l'apprentissage du système. L'université de Karlsruhe (Allemagne) effectue un travail analogue pour réaliser un corpus en allemand. La conception et la réalisation du système de reconnaissance sont confiées à la société belge Lernout & Hauspie.

Cet article est organisé de la façon suivante. La section 2 présente quelques corpus équivalents au notre. La section 3 présente une description du langage de commande du système. La section 4 décrit les sources de bruits identifiées. La section 5 décrit le protocole d'enregistrement. La section 6 donne notre méthodologie pour étiqueter facilement le corpus. La section 7 s'intéresse à la certification des données. Enfin la section 8 présente une conclusion de l'article.

2. ÉTAT DE L'ART

Il n'existe pas de corpus français enregistré dans une voiture. Parmi les corpus enregistrés en présence de bruit et à l'aide d'un microphone éloigné du locuteur, on peut citer "Far Field Microphone TIMIT" qui est la version de TIMIT – parole lue en anglais – enregistrée à partir du microphone secondaire qui captait aussi bien la parole que le bruit de la ventilation riche en basse fréquence. Parmi les corpus incluant de la parole bruitée en français, on peut citer BDBRUIT qui contient des enregistrements de 10 locuteurs prononçant 30 phrases, 30 chiffres et quelques logatomes en présence de bruit blanc ou dans une ambiance de cocktail. Notre contribution dans le projet VODIS a permis d'enregistrer 200 personnes à l'aide de 5 microphones dans des ambiances de bruits automobiles différents. L'ensemble occupe 13 CDroms.

3. DESCRIPTION DU CORPUS ENREGISTRÉ

Le vocabulaire de commandes est constitué actuellement de 140 mots-clés. Des recherches en vue d'intégrer un langage de requêtes spontanées sont en cours. Les 140 mots sont basés sur les commandes du système *Travel Pilot* RG05 de la société Bosch/Blaupunkt.

couper le son	remettre le son
mode automatique	mode manuel
aigus ; graves	volume, radio
cassette, télephone	CD, navigation
OK	c'est bon
mettre + *ACRONYM*	appeler + *NAME*
page précédente	correction
choisir destination	commencer guidage
afficher la carte	répéter
guidage vocal	guidage silencieux
menu principal	C cédille
accent circonflexe	*les lettres de alphabet*
les chiffres de 0 à 16	20 ; 21 ; 30 ; 31 ... 1000

TAB. 1 – *Une partie des différents prompts*

Le vocabulaire comprend également l'épellation des lettres afin de prononcer les noms de villes destinés à la commande du système de navigation et l'ensemble des chiffres pour la numérotation téléphonique (cf. table 1). Voici quelques exemples :

– *mettre RTL, choisir la destination, ville, N A N C Y, rue, L E C L E R C, afficher la carte, couper le son, téléphone, décrocher, appeler le 03 83 59 20 79, raccrocher, remettre le son.*

Cet ensemble de requêtes orales correspond à un ensemble de requêtes accessibles à tout instant par des menus déroulants du système *Travel Pilot*. L'utilisation de la parole permet un accès plus facile au système par l'utilisation de raccourcis que sont les requêtes orales.

L'enregistrement du corpus à bord de la voiture équipée du système de navigation se fait à deux personnes : le conducteur qui est aussi le locuteur et un copilote, responsable de la prise de son et qui présente des prompts au conducteur. On distingue ainsi 3 types de requêtes suivant l'interaction entre le copilote et le conducteur :

– les requêtes promptées visuellement ou sonores ;

– les requêtes semi-promptées dans lesquelles le locuteur possède une part de décision, comme : *appelez un numéro de téléphone de votre choix* ;

– et les requêtes libres principalement destinées au système de navigation dans lesquelles le locuteur, après avoir été mis en situation, demande quelque chose au système. Ainsi la requête énoncée par le copilote comme – *demander au système de navigation de vous orienter vers une rue de votre choix.* – donne naissance à la requête prononcée par le conducteur – *chercher rue Mangin.* – ou – *rechercher à Dompierre le 4 place de la mairie.*

Tous les prompts sonores (PS) et visuels (PV) sont crées aléatoirement avant le début du trajet et constituent un fichier qui sera utilisé pour l'étiquetage. Chaque conducteur prononce un ensemble de 180 phrases réparties dans les catégories suivantes :

- 73 mots clés de commande (PS) ;
- 26 lettres par suites de 4 (PV) ;
- certains signes diacritiques (PS) ;
- les chiffres par groupe de 4 et 3 (PV) ;
- des nombres (PS) ;
- des numéros de téléphone (PV) ;
- 5 phrases phonétiquement équilibrées (PS).

Ces dernières phrases permettentt une modélisation en phonèmes pour la reconnaissance des noms de stations de radios, ou la reconnaissance de noms et de prénoms de personnes enregistrés dans un répertoire téléphonique et plus tard la reconnaissance de requêtes spontanées.

4. DESCRIPTION DE SOURCES DE BRUITS ENREGISTRÉES

Le bruit à l'intérieur d'une voiture dépend de paramètres pilotés et subis. Parmi les facteurs extérieurs subis, l'état de la route joue une place importante. L'utilisation d'accessoires bruyants peut aussi influer sur le signal enregistré. Nous avons choisi d'enregistrer et d'étiqueter une partie des bruits caractéristiques à l'intérieur d'une voiture. Six classes de bruit peuvent affecter le signal enregistré ont été déterminées. Chaque classe de bruit est représentée par un champ dans l'entête du fichier son. La table 4 répertorie une partie des champs de l'entête.

Ainsi, l'entête du fichier son donnée dans la table 3 indique que le mot *quatre-vingt onze* a été prononcé fenêtre fermée, sans ventilation, sur une autoroute sèche et à une vitesse comprise en 90 et 120 km/h.

5. PROTOCOLE D'ENREGISTREMENT

Afin d'obtenir une séparation de sources, trois microphones espacés de 30 cm sont situés à proximité du pare-soleil

nom	sens	valeur possibles
radio	radio	off, on mono, on stereo
crosstalk	conversations voisines	none, passengers front passenger
weather	météo du moment	normal, rain hail, storm
road type	type de route	normal, potholes coble-stone pavement
road surroundings	abords	normal, alley, tunnel, construction site
noises other	description libre du bruit	wiper, flashing, etc.

TAB. 2 – *description des sources de bruit*

du conducteur. Leurs signaux sont enregistrés sur un DAT à 8 canaux avec une bande passante de 10 kHz. Le signal de l'auto-radio est enregistré sur une piste séparée afin de tester les algorithmes de valorisation de la parole du conducteur *speech enhancement*. Le conducteur porte un microphone de proximité de type Shure SM 10A qui sert de référence pour l'étiquetage. Le sixième canal est à la disposition du copilote qui enregistre un index permettant de retrouver chaque phrase prononcée et sélectionnée pendant l'enregistrement continu du trajet. Le rapport moyen signal sur bruit varie de 50 dB pour le microphone de proximité à moins de 20 dB pour les microphones situés sur le plafonnier.

Avant chaque enregistrement un scénario est défini qui fixe les conditions de prises de son : la vitesse, l'ouverture de la fenêtre, la mise en marche de la ventilation et de la radio, etc.

L'enregistrement de 200 conducteurs (100 hommes, 100 femmes) se fait dans une voiture en situation réelle de circulation. Trois voitures ont été utilisées dans cette campagne d'enregistrement. Une prise de son dure 45 minutes par personne. Les requêtes au système ainsi que les mots du vocabulaire sont répétés après un prompt sonore ou visuel. Les bruits variant au cours de l'enregistrement (exp : pluie, état de la route) sont notés régulièrement par l'opérateur. Des pauses régulières sont faites afin de garder une élocution naturelle. A la descente de voiture, le scénario contenant l'ensemble des prompts est modifié pour tenir compte des derniers changements. Il est aussi complété par les annotations sur les bruits rencontrés (cf. 4).

Les données du DAT obtenues sont démultiplexées. Le sixième canal permet de créer un fichier d'index contenant l'adresse (en millisecondes) de chaque début de parole. Les signaux des quatre microphones (et de la musique si nécessaire) sont filtrés et sous-échantillonés à 11 025 Hz.

6. ÉTIQUETAGE DES PHRASES

L'étiquetage porte sur le canal réservé au microphone de proximité porté par le conducteur. Chaque phrase se retrouve dans 4 ou 5 fichiers suivant le canal enregistré.

Un algorithme de détection parole/non-parole [4] recherche le début de l'élocution à partir de l'index enregistré dans la voiture sur le sixième canal. Cette séquence est étiquetée par défaut à l'aide du prompt correspondant dans le fichier scénario (cf. § 5). Les limites placées sont vérifiées visuellement grâce à un éditeur de signal développé dans notre équipe[2]. L'opérateur peut avancer à sa guise dans le fichier des index ou des prompts pour maintenir le synchronisme entre les deux listes.

Les limites placées sont utilisées pour générer les fichiers sons contenant 1/2 seconde de silence avant et après la parole. Chaque mot du vocabulaire, et ceci pour chacun des 4 microphones, constitue un fichier son possédant une entête au format NIST [3] qui est crée automatiquement grâce au scénario établi préalablement. Cette entête contient également des informations sur le locuteur et les conditions d'enregistrement. Ces données ainsi faites sont alors vérifiées et validées par un système de reconnaissance utilisant des modèles de Markov qui essaie de reconnaître n'importe quelle suite de mots du vocabulaire dans chaque fichier.

```
NIST_1A
   1024
author -s14 INRIA Lorraine
sample_rate -i 11025
sample_n_bytes -i 2
channel_count -i 1
microphone -s11 SHURE MS10A
speaker_id -s3 019
speaker_age -i 25
speaker_dialect -s4 none
recording_date -s8 01-04-97
microphone_position -s10 close-talk
speaking_mode -s6 prompt
driving_situat. -s20 long-dist. trafic
weather -s3 dry
road_type -s7 highway
road_condition -s3 dry
fan -s3 off
window -s6 closed
speaker_gender -s4 male
speaker_school_town -s7 unknown
noises_other -s4 none
road_surroundings -s6 normal
crosstalk -s4 none
car_type -s11 Peugeot 406
radio -s3 off
transliteration -s17 quatre-vingt-onze
driving_speed -s5 90-120
end_head
```

TAB. 3 – *Extrait de l'entête du fichier son contenant une prononciation du mot quatre-vingt-onze au travers du microphone de proximité*

7. CERTIFICATION

Afin de valider l'étiquetage, nous effectuons une reconnaissance automatique de chaque élément du corpus. Le signal est paramétré à l'aide de 8 coefficients cepstraux et des 16 coefficients de régression du premier et second ordre. Chaque mot du vocabulaire est représenté par un modèle de Markov du second-ordre (HMM2) [5]. L'apprentissage et la reconnaissance des modèles se fait en

utilisant la parole des 200 conducteurs provenant du microphone de proximité afin de s'affranchir au mieux des bruits du véhicule et d'utiliser un signal de bon rapport signal / bruit. Le nombre d'états des HMM2 varie suivant le nombre de phonèmes du groupe de mots : 5 états pour les plus petits comme la lettre *A*, 10 états pour *couper le son* ou encore 21 comme dans *Montrer la destination*. Chaque état représente une loi normale avec une matrice de covariances complète. Ces fichiers contiennent 1/2 seconde de silence en début et fin. Ce silence est incoporé dans la modélisation des unités initales et finales. Les phrases phonétiquement équilibrées ne sont pas utilisées au cours de cette étude. Il s'agit d'une reconnaissance globale multilocuteur dans laquelle le corpus de test est égal au corpus d'apprentissage. Les réponses du reconnaisseur sont comparées à l'étiquetage défini dans la section 6. Cette phase de reconnaissance permet de déceler l'existence de fichiers mal étiquetés qui ne correspondent pas au mot réellement prononcé ou qui contiennent un bruit anormal gênant la reconnaisance. L'opérateur peut alors vérifier, corriger ou supprimer ces données erronées. Cette démarche s'apparente à celle développée par [6] pour étiqueter de grosses bases de données. Elle permet d'effectuer rigoureusement une vérification de l'étiquettage correct du corpus, gage d'un bon apprentissage par la suite.

nature du mot	Tx. correct(%)	Tx. reco.(%)
mots clés	97	95
chiffres de 0 à 16 20, 30, etc.	92	91
lettres	95	91
nombres	91	81

TAB. 4 – *Taux de reconnaissance sur les mots clés, chiffres , nombres et lettres*

Pour la reconnaissance, nous utilisons une grammaire de suites de mots de longueur quelconque. La table 4 donne les taux de reconnaissance sur les 73 mots clés de commande du système de navigation incluant les commandes pour l'auto radio et le téléphone, ainsi que les 29 chiffres et les nombres. Le taux correct (tx. correct) donne le pourcentage de mots reconnus par le système sans aucune erreur d'omission ni de substitution mais en autorisant les insertions. Le taux de reconnaissance (tx. reco.) donne le pourcentage de mots reconnus sans substitution, insertion et ommission. Ces taux montrent qu'en moyenne 10 % des fichiers doivent être individuellement re-vérifiés. Les résultats obtenus pour les nombres s'expliquent par le fait que l'apprentissage s'est fait en mode isolé à partir des racines de prononciation des nombres. Les phénomènes de coarticulation ne sont pas appris. Les différences entre les taux corrects et taux de reconnaissance s'expliquent par l'absence de contraintes linguistiques représentées par des bigrammes. Toutes les suites de modèles sont autorisées et de nombreuses insertions peuvent apparaître. Ceci pénalise les résultats de reconnaissance et rend la certification des fichiers plus difficile.

8. CONCLUSIONS

L'enregistrement d'un corpus oral destiné à l'apprentissage d'un système embarqué de reconnaissance pose des problèmes spécifiques. L'interaction de l'usager avec le système doit être prise en compte dés le début afin de déterminer le vocabulaire du système. Le protocole d'enregistrement doit être soigneusement spécifié afin de mettre chaque locuteur le plus rapidement et le plus facilement en situation. La reconnaissance de la parole à l'intérieur d'une voiture doit tenir compte de plusieurs problèmes spécifiques comme la séparation des sources ou la présence de bruits. Nous avons spécifié et enregistré un corpus multicanaux. Chaque fichier a été annoté à l'aide des bruits caractéristiques audibles dans une voiture. En dernier lieu, l'automatisation partielle des différentes étapes d'étiquetage et de certification des données permet de diminuer le temps de traitement et d'éliminer toute erreur humaine d'étiquetage. Ce corpus est maintenant disponible pour la communauté participant au projet VODIS et sera distribué à terme dans le monde de la recherche.

RÉFÉRENCES

[1] M. Drbecker. Small microphone arrays with optimized directivity for speech enhancement. In *Proc. EuroSpeech'97*, pages S. 327 – 330, Rhodes, September 1997.

[2] D. Fohr and Y. Laprie. Snorri : An Interactive Tool for Speech Analysis. In *Proceedings EUROSPEECH 89 (European Conference on Speech Communication and Technology)*, Paris, September 1989.

[3] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallet, and N.L. Dahlgren. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM , 1993.

[4] J-C. Junqua, B. Mak, and B. Reaves. A Robust Algorithm for Word Boundary Detection in the Presence of Noise. *IEEE Transactions on Speech and Audio Processing*, 2(3):406 – 412, july 1994.

[5] J.-F. Mari, J.-P. Haton, and A. Kriouile. Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5:22 – 25, January 1997.

[6] Odile Mella and Dominique Fohr. Semi-automatic phonetic labelling of large corpora. In *EUROSPEECH*, volume 4, Rhodes, Greece, September 1997.

[7] P Pollak, P. Sovka, and J. Uhlir. Noise suppression system for a car. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1073 – 1076, 1993.

[8] O. Siohan, Y. Gong, and J.-P. Haton. Noise Adaptation using Linear Regression for Cont inuous Noisy Speech Recognition. In *Proceedings 4th European Conference on Speech Comm unication and Technology*, volume 1, pages 465–468, Madrid (Spain), September 1995.

EMACOP : Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole

Vaufreydaz Dominique, Akbar Mohammad, Caelen Jean, Serignat Jean-François

Laboratoire CLIPS (Communication Langagière et Interaction Personne-Système)
Université Joseph Fourrier, 38041 Grenoble cedex 9 France
Tél.: +33 4 76 51 46 27
e-mail: [Dominique.Vaufreydaz, Mohammad.Akbar, Jean.Caelen, Jean-Francois.Serignat]@imag.fr

ABSTRACT

This paper presents an Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 95 and windows NT. It provides graphical interface functions to process the signal in order to compute acoustical parameters used in speech analysis, recognition and synthesis. Entries' extraction from main database is also provided from other sources. The new way of programming through the Microsoft COM (Component Object Model), which gives encapsulation/aggregation mechanisms and independence for each module, is used to increase reliability, flexibility and maintainability.

1. INTRODUCTION

La nécessité de disposer des bases de données sonores a émergé il y a une dizaine d'années, sous la poussée des méthodes de reconnaissance de la parole fondées sur l'apprentissage. Dans le milieu des années 80, des travaux ont été coordonnés en France ([Car84], [Des86]) autour de BDSONS, et prolongés dans des actions européennes (projets SAM I et SAM II). A l'heure actuelle, le besoin de données sonores reste encore une préoccupation essentielle.

Dans le même temps, les ordinateurs personnels (PCs) se sont répandus et ont atteint des performances intéressantes pour le traitement de la parole ; leur rapport prix/puissance n'a pas d'équivalent dans le monde des stations de travail Unix. Il est donc intéressant de les utiliser pour réaliser toutes les phases nécessaires à l'acquisition de signaux de parole et à leur stockage pour obtenir des environnements complets de traitement de la parole.

Dans le monde PC, de nombreux logiciels de portée générale existent et il aurait été plus rapide de coupler des utilitaires de bases de données et des logiciels de traitement de signal ([Akb97]) existants, pour construire un ensemble applicatif couvrant nos besoins. Mais l'inconvénient de cette solution réside dans le coût élevé de l'exploitation. En effet, ces logiciels sont souvent chers et leur utilisation, n'est pas, dans la plupart des cas, orientée vers un problème précis. En outre, cela aurait reporté une grande partie de la gestion (en particulier des fichiers) sur l'utilisateur. Enfin cela n'aurait pas permis de contrôler aisément et sans programmation les paramètres d'une session d'enregistrement. C'est pourquoi, malgré l'existence de logiciels comme EUROPEC [Zei91], développé sous DOS et utilisant une carte d'acquisition spécifique, nous avons développé une version plus étendue sous Windows 95 et Windows NT ne mettant en œuvre que du matériel standard : EMACOP (Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole).

La plupart du temps, les campagnes d'enregistrement mobilisent d'importantes ressources humaines pour guider ou assister les locuteurs dans leur tâche de diction, pour organiser l'enregistrement, pour préparer les scénarios et les données, etc. Il faut pouvoir contrôler les différents scénarios pour varier les conditions de capture : la lecture d'un texte ou d'une suite de mots ou de mots isolés, la répétition après écoute d'une phrase, le dialogue en réponse à des questions, etc. Les méthodes d'acquisition rigoureusement contrôlées sont donc lourdes. Elles nécessitent même un surcoût de travail, si l'acquisition numérique n 'est pas effectuée en direct, car il faut convertir et échantillonner les bandes analogiques et étiqueter les items sonores.

C'est pourquoi le développement d'un utilitaire de gestion et d'acquisition de grands corpus sur un matériel standard, nous semble d'un grand bénéfice à condition que la convivialité et la simplicité d'utilisation permettent à des non-spécialistes de maîtriser l'application.

2. L'APPLICATION

L'ensemble du logiciel se présente comme deux applications distinctes qui communiquent via un réseau grâce au protocole TCP/IP[1]. Elles sont toutes deux construites sur un modèle de composant (COM de Microsoft) qui les rend hautement modifiables et en facilite la maintenance : cela permet notamment l'accès à de multiples interfaces pour le même module, la recompilation d'un module sans retoucher aux autres, et il est possible de répartir de manière transparente

[1] Transfer Control Protocol / Internet Protocol

l'exécution des différents composants sur différentes machines.

La première application est le *serveur* qui permet la définition de la base de données et qui n'est exécutée que sur une seule machine. Elle fonctionne comme un serveur qui accepte les communications sur un "well known port" (qui peut être défini à chaque lancement) avec plusieurs clients.

La seconde application est le *client*. Elle a en charge toutes les fonctionnalités de présentation des items et d'acquisition du signal. Dans une optique d'exportation de savoir-faire du serveur vers le client, cela permet d'utiliser au maximum la machine locale et de décharger le serveur de tout traitement, autre que celui des données.

La communication entre le serveur et ses clients, se fait au moyen de requêtes, en mode connecté mais, dans le but de palier les problèmes réseaux, l'application cliente charge le maximum de données à la fois et peut, lors d'une coupure momentanée des communications, travailler avec les données qu'elle possède en local. Enfin si la durée de l'interruption dépasse un certain délai d'attente raisonnable pour l'utilisateur, il est possible d'envoyer plus tard au serveur le résultat de la session d'enregistrement en cours. Il sera alors possible de reprendre la session de travail là où l'on s'est arrêté auparavant. Un autre avantage de cette architecture semi-décentralisée est qu'elle peut supporter un nombre quelconque de clients dans la mesure où les transferts de données client-serveur peuvent être entrelacés avec l'enregistrement du signal sur le client.

Du fait qu'il existe deux applications avec des fonctionnalités différentes et que les niveaux de connaissance pour les gérer et les paramétrer sont différents, nous devons faire une distinction au niveau des compétences. Nous appelons **superviseur** la personne qui a en charge la gestion du serveur et de toutes les données qu'il contient, et **locuteur** toute personne qui utilise le logiciel client pour réaliser l'acquisition. Ces deux personnes peuvent évidemment être la même.

2.1. *Le serveur*

La définition des données

La définition des données se fait à l'aide d'un utilitaire qui connaît plusieurs types d'entités : les corpus, leurs éléments, les locuteurs, les scénarios d'acquisition. Ces entités constituent les éléments de la base de données.

Après le lancement du serveur, une boîte de dialogue permet de choisir les modifications que l'on veut apporter à la base. L'interface se compose d'un enchaînement de fenêtres permettant la création des diverses entités et des liens entre celles-ci. Elle permet, non seulement l'entrée, mais aussi le contrôle de certaines données, qui sont, soit mises en forme automatiquement lors de la saisie, soit limitées, avant saisie, à un sous ensemble de valeurs (définitions standards) pour limiter les erreurs de frappe. Une meilleure cohérence du contenu de la base de données est ainsi garantie.

Le superviseur peut aussi, à tout moment, modifier les définitions standards en ce qui concerne les langues, les caractéristiques linguistiques et les types de corpus. Les définitions de base sont tirées de la norme ESPRIT-SAM [Tom91]. Toutes ces données sont de nature à pouvoir être utilisées dans le traitement automatique de la parole. Ainsi, pour un corpus, on définit son type (par ex. série de chiffres, de phrases, de textes, etc.), la langue et enfin la liste de ses éléments. Un locuteur est défini par son nom, son prénom, sa langue maternelle, sa caractéristique linguistique (par exemple parisien). Lors de l'exportation un code locuteur anonyme remplace le couple nom-prénom. A ceci s'ajoutent des caractéristiques physiques permettant de faire une différenciation plus précise des locuteurs : son âge (au moment de l'enregistrement), sa taille, son poids, son sexe et enfin s'il est fumeur ou non.

A l'aide de toutes ces informations, il est possible de rattacher divers renseignements aux signaux de parole enregistrés. La figure 1 schématise les données au sein de la base, les relations entre celles-ci et l'interface qui est présentée au locuteur pendant la tâche d'acquisition. Chaque locuteur connecté "voit" ainsi la partie de la base de données qui lui est affectée.

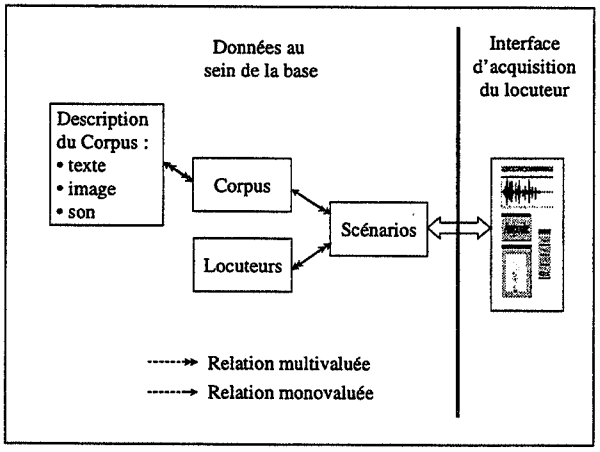


Figure 1 : Les données de la base gérées par le superviseur et l'interface du locuteur. Le superviseur définit les scénarios et gère la base de données (corpus, locuteurs et leurs descriptions).

Comme le montre la figure 1, la partie la plus importante du travail du superviseur est la définition des scénarios d'acquisition. Ces scénarios modélisent les plans de déroulement des sessions d'acquisition et permettent de définir les modalités d'affichage des stimuli. On notera que plusieurs scénarios peuvent être affectés à un locuteur, qu'à un scénario ne correspond qu'un seul corpus, lui même constitué d'un ensemble non limité de données.

La description d'un scénario est réalisée au moyen d'une interface qui propose au superviseur de faire des choix entre diverses alternatives et de définir plusieurs valeurs.

C'est ainsi qu'il détermine de quelle manière vont être présentés au locuteur les éléments à prononcer. Trois types de présentation sont disponibles. L'audition de ce qui doit être dit, l'affichage d'un texte ou l'apparition d'une image. Ces modes de présentation peuvent être utilisés de façon complémentaire. A ceci s'ajoute un temps de présentation qui détermine la durée de l'affichage des données en millisecondes. Enfin le superviseur indique le nombre de fois que doivent être répétés les items.

A ce niveau, il est possible de tenir compte des compétences du locuteur pour modifier le scénario :

- S'il ne possède aucune connaissance en matière de signal, on décide de lui faire énoncer les corpus sans qu'il n'ait aucune action sur ce qu'il produit,
- Il peut écouter ce qu'il a enregistré après chaque item, et juger par lui-même s'il le signal acquis est correct et s'il doit être archivé ou non,
- Il peut visualiser en temps réel ce qu'il enregistre pour en contrôler la pertinence,
- Il peut lui être alloué le droit de réaliser, pour chaque enregistrement, des modifications sur le signal correspondant à "couper" ou "modifier", pour éliminer les silences inutiles.

Le traitement des données

Le superviseur peut connaître, grâce à l'historique, toutes les actions exécutées par les locuteurs durant leurs sessions d'enregistrement. Cela lui permet de contrôler leur travail, s'il n'était pas présent ou si cela s'est fait *via* le réseau par exemple. Il trouve dans cet historique les annulations, les répétitions et les modifications que le locuteur a effectuées. Il a aussi accès à la connaissance du temps de réponse de ce dernier. Ce temps est calculé comme la différence entre l'instant d'affichage (ou de fin d'écoute) et l'instant de détection du début de parole.

Le superviseur peut faire des retouches de signal *a posteriori*. Il peut aussi vérifier ce qu'a fait et dit le locuteur. Ainsi les hésitations ou les substitutions de mots peuvent être traitées et/ou éliminées (ce système de correction a été utilisé pour la vérification de la base de données BREF [Lam91] par exemple). On possède ainsi à la fin de l'acquisition, une version des données épurées en vue de l'apprentissage ou de l'étiquetage.

Le superviseur peut exporter les signaux en vue de leur exploitation par d'autres logiciels. L'exportation se fait grâce à l'affichage de différentes arborescences (triées par locuteurs, par corpus ou par éléments) qui permettent aisément de choisir les sons à exporter selon les critères souhaités. Les formats SAM et ECHO [Kab96] sont par exemple supportés. Il est facile de définir son propre format d'exportation en créant un composant ayant la même interface que celui qui est livré avec l'application.

2.2. *LES CLIENTS*

Initialisation de l'application cliente

Après avoir défini entièrement toutes les modalités de présentation, un locuteur peut enregistrer. Il commence par s'identifier auprès du serveur dans le double but de lui attribuer une connexion et de lui délivrer les scénarios qui lui sont affectés. Il peut en choisir un, le mener à son terme ou l'arrêter en cours de route, le reprendre ou l'abandonner. Le locuteur peut sélectionner, dans un ordre quelconque, les scénarios d'acquisition qui lui sont demandés d'effectuer.

Un outil graphique de calibrage lui permet de déterminer les seuils de détection de sa voix par rapport à l'environnement sonore dans lequel il réalise son enregistrement. Cet outil lui permet de calculer automatiquement ces seuils et de les ajuster ensuite pour obtenir de meilleurs résultats. Ceci permet de ne pas être dépendant d'une interface avec des boutons pour déclencher l'enregistrement. On limite ainsi les effets de coupure et la présence de bruits, en début ou fin de signal, ce qui donne une meilleure qualité au signal produit même lorsque le locuteur n'a aucun droit de retouche sur celui-ci.

Interface générale pendant la phase d'acquisition

L'interface destinée au locuteur est présentée sur la fig. 2.

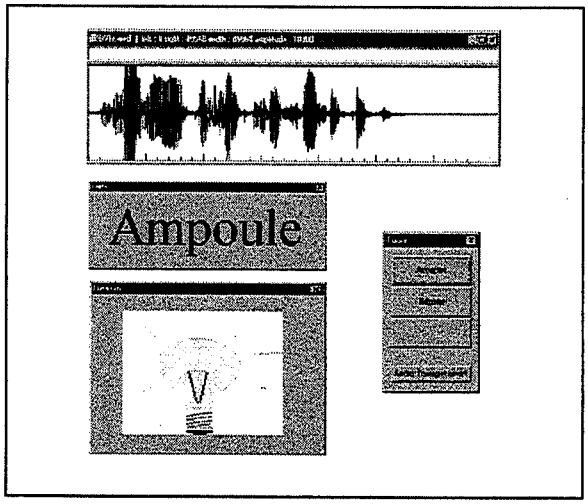


Figure 2 : Interface complète destinée au locuteur.

Elle est ici dans l'état où l'on présente un mot isolé *ampoule* et une image au locuteur. Dans cette situation, le locuteur a enregistré le mot et peut observer l'onde sonore acquise. Il peut éditer le signal et le modifier, puis il peut valider ou invalider l'enregistrement avec les boutons de la fenêtre de contrôle (ici l'enregistrement ne correspond manifestement pas à l'énoncé du mot *ampoule*).

Cette interface est composée de plusieurs fenêtres indépendantes, au maximum quatre selon les droits du locuteur (si le locuteur possède la visualisation et/ou

l'édition du signal produit et le mode de présentation par image et texte). De plus, il possède la visualisation du signal qu'il produit. La modularité de l'interface est totale puisque le locuteur peut placer à sa guise chaque fenêtre là où il le désire sur son écran.

La première qui se présente est celle contenant le menu de sélection. Elle lui permet dans un premier temps de lancer le scénario ou de revenir en arrière. Lors de la phase d'acquisition à proprement parler, un plus grand choix lui est proposé. Il peut accepter le signal qu'il vient de prononcer et passer au suivant. Il peut écouter ou voir le signal qu'il a produit (s'il en a le droit). Enfin il peut arrêter l'acquisition pour la reprendre plus tard.

La seconde fenêtre est celle qui présente le texte. Celui-ci est présenté dans une fonte de taille variable. L'affichage compte au maximum 4 lignes de 50 caractères (pour des raisons de lisibilité). Si l'élément dépasse cet affichage, des barres de défilement ("scrolling") apparaissent. Dans le cas où un temps de présentation est défini, ces barres sont remplacées par un défilement qui se déroule automatiquement et proportionnellement au texte.

La troisième fenêtre est celle qui autorise l'affichage des images. Celles-ci peuvent être de natures différentes (dessin ou image). Cela donne à EMACOP une portée d'utilisation en psycholinguistique par exemple : il peut été utilisé pour l'étude de normes associatives pour des images ou des sons, comme il en existe pour les mots [Mei98].

La dernière fenêtre est celle qui affiche, en temps réel, le signal pendant la phase d'acquisition et qui permet l'édition de celui-ci lorsque cela est possible. Elle propose un affichage sous forme d'onde (cf. figure 2) ou sous forme de nuage de points. La précision (nombre d'échantillons par pixel) peut être modifiée. Le locuteur peut, à l'aide d'un menu contextuel et d'une sélection à la souris, écouter tout ou partie du signal et réaliser des "couper" pour l'épurer.

2.3. *VALIDATION*

Le logiciel a été validé en situation de fonctionnement nominal sur une grande base de données. Les échanges à travers le réseau sont en cours de test.

3. CONCLUSION ET PERSPECTIVES

L'ensemble d'acquisition qui est décrit ci-dessus permet une très grande précision au niveau de la définition des différentes entités utilisées dans la base de données. Ainsi on peut être très rigoureux dans la définition des éléments et obtenir pour chaque scénario, une bonne adéquation entre le type de signaux désirés et la présentation des items. De plus, l'utilisation du réseau permet de faire des acquisitions en parallèle et donc facilite les grandes campagnes d'acquisition de corpus, éventuellement en parallèle sur différents sites. Enfin EMACOP autorise l'acquisition à grande distance (Internet) des corpus et permet enregistrer une plus grande diversité de personnes.

L'évaluation complète d'EMACOP est actuellement en cours et doit permettre de vérifier que l'on atteint un très bon contrôle de la prosodie, du débit et de la qualité de signal enregistré.

Enfin pour répondre aux attentes de nombreux utilisateurs, l'enregistrement en stéréo sera mis en place. Cela rendra possible l'enregistrement de dialogues, de fonds sonores mais aussi de signal provenant, par exemple d'un laryngophone.

BIBLIOGRAPHIE

[Akb97] Akbar M. (1997), « WaveEdit, An Interactive Speech Processing Environment for Microsoft Windows Platform ». EuroSpeech'97. Vol. 2, pp. 677-680.

[Car84] Carre R., Descout R., Mariani M., Rossi M. (1984), « The French Language Database : Defining, Planning and Recording a large Database ». Proc. IEEE ICASSP, San Diego, USA, 42.11.

[Des86] Descout R., Serignat J.F., Cervantes O., Carre R., (1986), « BDSONS : Une base de données des sons du français », 12th International Congress on Acoustic, Toronto, Canada, Vol. A, pp. 4-7.

[Kab96] Kabré, H. (1996), « ECHO: a Speech Input/Output system for the design of Interactive Speech-based Applications », American Voice Input/Output Systems, San-Jose, USA, pp. 179-188.

[Lam91] Lamel L.F., Gauvain J.L., Eskenazi M. (1991), « BREF, a Large Vocabulary Spoken Corpus For French », EuroSpeech'91.

[Mei98] Meiye J.P., Rouillard J., Vaufreydaz D. (1998), WebCompletion – Protocole de Normes Associatives sur Internet. Ecole Sémantique CNRS Caen 98. http://www-multicom.imag.fr/Cognition/Normes

[Tom91] Tomlison M.J. (1991), Guide to Database Generation – Recording Protocol, Final Version. SAM-RSRE-015, Marlvern, Angleterre.

[Zei91] Zeiliger J., Serignat J.F. (1991), Europec software (v4.1), User's Guide Release 4.1. SAM-ICP-045, Grenoble, France.

Langue

Identification des langues par discrimination automatique des systèmes vocaliques

F. Pellegrino, A. Galinier

IRIT
118, route de Narbonne – F-31062 Toulouse Cedex - FRANCE
Tél.: ++33 5 61 55 60 55 - Fax: ++33 5 61 55 62 58
e-mail: {pellegri, galinier}@irit.fr - http://www.irit.fr/~Francois.Pellegrino

ABSTRACT

This paper deals with our research on vowel system modelling in an Automatic Language Identification (ALI) purpose. The study of vowel systems shows that they carry an important part of the language characteristics and that taking advantage of this knowledge is promising. We propose an alternative modelling to the standard acoustic-phonetic decoding currently used as front-end in the ALI systems: we model each language vocalic system as a Gaussian mixture that is estimated out from automatically detected vowels. With 5 language from the OGI MLTS corpus and in a close set identification task, we reach 62.7 % of correct identification with 45 second duration utterances for the male speakers. Taking into account that we use only the vowel information (less than 15 seconds per utterance) and no language modelling, these results offer many perspectives.

1. INTRODUCTION

L'Identification Automatique des Langues (IAL) émerge comme un enjeu majeur du Traitement Automatique de la Parole. Alors que des interfaces homme-machine (IHM) performantes et fiables sont maintenant disponibles, le développement des communications mondiales et le cadre multi-ethnique de la société de cette fin de $XX^{ème}$ siècle sont à l'origine d'une demande croissante pour des systèmes multilingues. Cet obstacle linguistique demeurera tant que les systèmes d'IAL n'atteindront pas des performances et une fiabilité suffisantes pour ne plus être le maillon faible des IHM.

Développer un système intrinsèquement multilingue est difficile [4], et la prise en compte de plusieurs langues se fait généralement au prix d'une baisse des performances dans chacune d'entre elles. Une manière rationnelle d'obtenir des systèmes multilingues est donc de développer un système dédié à l'IAL et chargé d'orienter le traitement vers un système de TAP spécifique à la langue reconnue.

La plupart des études récentes [2, 3, 11] montrent qu'un modèle de langage basé sur des statistiques de type n-gramme s'avère très efficace pour identifier des langues. La majeure partie des recherches menées se concentrent donc sur cette modélisation phonotactique. En contrepartie, la prise en compte des caractéristiques phonétiques et phonologiques en IAL a été étudiée de manière moins approfondie.

Notre travail a porté principalement sur cet aspect de la caractérisation automatique des langues pour plusieurs raisons. La majorité des systèmes requièrent une transcription phonétique des données. Étiqueter du signal de parole représente un investissement en temps et en expertise très important. Ainsi, étendre les capacités d'un système d'IAL à une nouvelle langue n'est pas aisé. D'autre part, il nous semble que le décodage acoustico-phonétique (DAP) utilisé par la plupart des systèmes n'est pas pleinement exploité mais uniquement utilisé comme projection d'un espace de paramètres acoustiques dans un espace symbolique discret. Le score d'identification de la langue est alors très peu influencé par cette phase qui constitue seulement un prétraitement à l'utilisation du modèle de langage.

Ces observations nous ont conduits à proposer une modélisation prenant en compte les systèmes phonologiques et phonétiques des langues à identifier. Nos travaux ont porté sur la caractérisation automatique des systèmes vocaliques (SV), afin de profiter de l'espace de représentation acoustique commun à toutes les voyelles. Ce choix a également été motivé par certaines observations faites sur la base UPSID [7, 10] : les 451 langues décrites partagent 307 systèmes vocaliques distincts (dont 271 qui sont spécifiques à une langue). Il est donc clair que, même si la description phonologique des systèmes vocaliques ne suffit pas à identifier toutes les langues, elle constitue un indice de discrimination important. Nous pensons que cette information, utilisée en corrélation avec des modèles de langages spécifiques, peut aboutir à un système d'IAL performant.

Le paragraphe suivant décrit le cadre théorique de notre approche. Le paragraphe 3 détaille plus précisément le système implanté et le paragraphe 4 décrit les expériences réalisées avec le corpus multilingue OGI.

2. LA MODELISATION DES SYSTEMES VOCALIQUES EN IAL

2.1 Synoptique du système

Le synoptique du système d'IAL étudié est illustré par la Figure 1. Le traitement appliqué au signal acoustique est basé sur l'utilisation d'une segmentation *a priori* basée sur l'algorithme de « divergence forward-backward » [1].

Un algorithme de détection de voyelles [9] permet d'extraire les segments vocaliques du signal. Chaque

segment est alors paramétré à l'issue d'une analyse cepstrale selon une échelle Mel. Cette phase de paramétrisation est totalement indépendante du langage et elle ne nécessite aucun étiquetage des données.

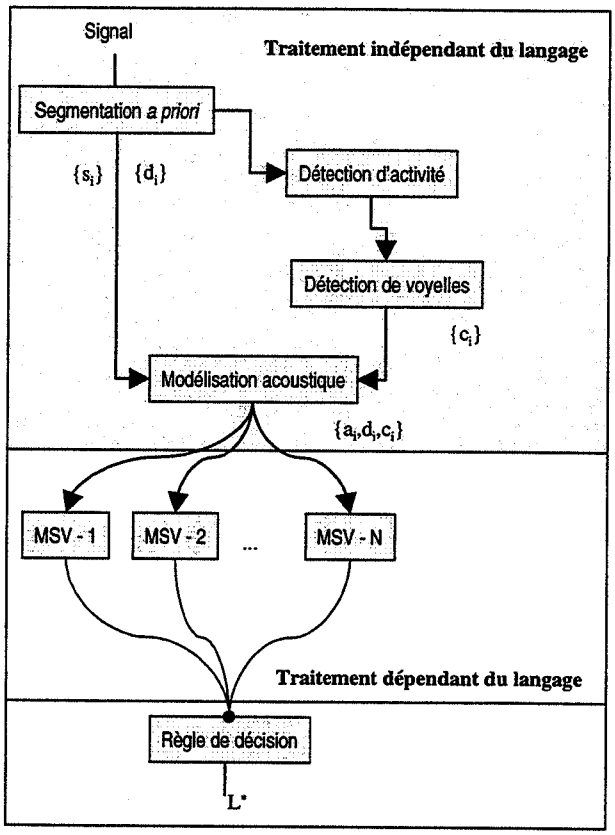


Figure 1. Schéma-bloc du système.

Un Modèle de Système Vocalique (MSV) est calculé sous forme de mélange de gaussiennes pour chaque langue traitée. Ces modèles sont appris avec toutes les voyelles détectées durant la phase d'apprentissage. En phase de reconnaissance, le score d'identification de chaque langue est donné par la vraisemblance des voyelles détectées dans la phrase testée par rapport au modèle. Un chemin symbolique sous la forme d'une suite Consonne/Voyelle est aussi fourni en vue d'une exploitation ultérieure dans un modèle de langage spécialement conçu.

2.2 Cadre statistique de l'étude

Les modèles employés sont dérivés des Modèles de Markov Cachés (MMC). De manière à étudier le pouvoir discriminant des systèmes vocaliques, nous avons modélisé toutes les consonnes avec un seul état, les autres états modélisant différentes qualités vocaliques.

Soit $L = \{L_1, L_2, ..., L_{NL}\}$ l'ensemble des N_L langues à identifier. Le problème est de déterminer la langue L^* dans laquelle une phrase inconnue est la plus probablement prononcée.

Soit $\Omega_i = \{C, V_1^i, V_2^i, ..., V_{ni}^i\}$ l'ensemble des états du modèle relatif à la langue i où C est l'état consonantique.

A l'issue du prétraitement acoustique, chaque segment est caractérisé par une concaténation de paramètres hétérogènes. Soient T le nombre de segments de la phrase traitée et $O = \{o_1, o_2, ..., o_T\}$ la séquence de vecteurs d'observation. Chaque vecteur o_i comprend le vecteur de coefficients cepstraux a_i, la durée du segment d_i, et une étiquette c_i, égale à 1 si le segment a été identifié comme une voyelle et 0 sinon. Nous notons $y_i = \{a_i, d_i\}$ et $o_i = \{y_i, c_i\}$.

Sachant la suite d'observation O, la langue la plus probable L^* est définie par :

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi i \in \Omega_i^T} \Pr(\Phi_i, L_i | O) \right] \quad (1)$$

où $\Phi_i = \{\varphi_1^i, \varphi_2^i, ..., \varphi_T^i\}$ est une séquence d'états du modèle de la langue i.

Cette définition peut-être reformulée en utilisant le théorème de Bayes :

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi i \in \Omega_i^T} \left(\Pr(O|\Phi_i, L_i).\Pr(\Phi_i|L_i) \right) \Pr(L_i) \right] \quad (2)$$

dans l'équation (2), on peut considérer que $\Pr(O|\Phi_i, L_i)$ est le terme de vraisemblance phonétique et que $\Pr(\Phi_i|L_i)$ traduit la vraisemblance phonotactique.

En se plaçant dans le cadre des hypothèses classiquement formulées avec les MMC, on considère que les segments sont indépendants les uns des autres conditionnellement aux états. Il nous semble que cette hypothèse est plus justifiable dans le cadre de segments homogènes de longueur variable que dans le cas classique de trames de longueur constante, mais il est évident qu'elle reste abusive. L'expression du modèle phonétique est alors :

$$\Pr(O|\Phi_i, L_i) = \prod_{k=1}^{T} \Pr(o_k|\varphi_k^i, L_i) = \prod_{k=1}^{T} \Pr((y_k, c_k)|\varphi_k^i, L_i) \quad (3)$$

qui peut encore s'exprimer sous la forme :

$$\Pr(O|\Phi_i, L_i) = \prod_{k=1}^{T} \Pr(y_k|c_k, \varphi_k^i, L_i) \Pr(c_k|\varphi_k^i, L_i) \quad (4)$$

La variable c_k étant supposée déterministe et conditionnée au symbole φ_k ($c_k = 0$ si $\varphi_k = C$ et $c_k = 1$ dans le cas contraire) on peut séparer la formule du modèle phonétique (4) en une partie consonantique et une partie vocalique. Si nous considérons de plus que le modèle consonantique est indépendant de la langue la vraisemblance du modèle phonétique se simplifie en :

$$\Pr(O|\Phi_i, L_i) = \prod_{ck=0} \Pr(y_k|\varphi_k^i = C). \prod_{ck=1} \Pr(y_k|\varphi_k^i, L_i) \quad (5)$$

En tenant compte des expressions (2) et (5), la langue la plus vraisemblable devient :

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi i \in \Omega_i^T} \left(\prod_{ck=1} \Pr(y_k|\varphi_k^i, L_i).\Pr(\Phi_i|L_i) \right) \Pr(L_i) \right] \quad (6)$$

Dans le cadre de notre étude, la probabilité d'occurrence des langues est supposée équiprobable. De plus, il nous est apparu intéressant de ne prendre en compte que les vraisemblances des systèmes vocaliques, sans considération phonotactique : Pr(ΦilLi) est considérée comme indépendante de la langue *i*. La langue la plus vraisemblable est alors donnée par :

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi i \in \Omega_i^T} \left(\prod_{\alpha t=1} \Pr(y_k | \varphi_k^i, L_i) \right) \right] \quad (7)$$

3 DESCRIPTION DU SYSTEME

3.1 Traitement acoustique

Cette étape du traitement a pour but de fournir une suite de segments étiquetés Consonne ou Voyelle ainsi que les paramètres associés. Trois prétraitements sont tout d'abord appliqués :

- L'algorithme de « Divergence Forward-backward » [1] produit une segmentation *a priori* sous forme de segments homogènes,
- Un détecteur d'activité vocale permet d'éliminer les phases de silences trop longs (pauses...),
- Un détecteur de voyelles [9] localise les segments vocaliques dans la phrase traitée.

Le traitement acoustique s'achève par le calcul de MFCC pour chaque voyelle détectée.

Prétraitement du signal

Les algorithmes de segmentation et de détection de voyelles ayant déjà été décrits [1 , 9], ils ne sont pas détaillés ici. La segmentation fait émerger des segments stables longs ainsi que des segments transitoires. La détection des voyelles est basée sur une analyse selon l'échelle Mel de la structure spectrale du signal. Cette technique ne requiert aucun apprentissage, et elle fonctionne sans connaissance *a priori* sur les données ; ce traitement est donc totalement indépendant de la langue.

Le détecteur d'activité vocale est conçu pour éliminer les longues pauses qui, ne portant pas d'informations phonétiques, ne seront pas traitées. L'algorithme effectue une analyse statistique simple du signal et un seuil T_a est calculé à partir de chaque segment : Soit z le signal acoustique et $S = \{s_1, s_2, ..., s_T\}$ la suite de segments. T_a est défini par :

$$T_a = \alpha . \min_{si}(\sigma_{Si}(z)) \quad (8)$$

où $\sigma_{Si}(z)$ est l'écart-type du signal dans le $i^{ème}$ segment. Le coefficient α a été expérimentalement fixé à 2,5.

Les segments présentant un écart type supérieur à T_a sont étiquetés comme étant de la parole, et dans le cas contraire ils sont identifiés comme étant du silence. La durée des segments intervient également puisque les silences d'une durée inférieure à 150 ms sont considérés comme significatifs (occlusions...) tandis que les segments plus longs sont assimilés à des zones de non-activité. Notons que si des segments voisins sont des silences, le seuil de durée s'applique à leur réunion.

Paramétrisation du signal

Chaque voyelle détectée est représentée par les 8 premiers coefficients cepstraux MFCC ainsi que leurs 8 dérivées. L'analyse cepstrale est effectuée sur une fenêtre de Hamming de 256 points centrée sur le milieu du segment. La durée du segment est ajoutée à ces paramètres cepstraux. Une soustraction cepstrale est appliquée en post-traitement de manière à atténuer l'influence du canal téléphonique. Pour chaque session d'enregistrement (même locuteur, même appel), un vecteur de MFCC moyen est calculé à partir de toutes les voyelles. Ce vecteur est alors soustrait des coefficients cepstraux de chacune des voyelles. Cette méthode est légèrement différente de la classique soustraction cepstrale où l'effet canal est estimé sur toute la phrase.

3.2 Modélisation des systèmes vocaliques

Les modèles de système vocalique (MSV) sont définis comme étant des MMC simplifiés, avec une topologie ergodique. Chaque état est modélisé par une densité de probabilité gaussienne (α, μ, Σ).

Soit $X = \{x_1, x_2, ..., x_N\}$ l'ensemble d'apprentissage et $\Pi = \{(\alpha_1, \mu_1, \Sigma_1), (\alpha_2, \mu_2, \Sigma_2), ..., (\alpha_n, \mu_n, \Sigma_n)\}$ l'ensemble de paramètres définissant le mélange de n gaussiennes dans un espace à p dimensions. Le modèle qui maximise la vraisemblance totale de l'ensemble d'apprentissage est :

$$\Pi^* = \arg\max_{\Pi} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{n} \frac{\alpha_k}{(2\pi)^{n/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] \right\}$$

où α_k est le poids de la $k^{ème}$ loi gaussienne.

S. Nowlan [8] a montré que sous les hypothèses :

- La somme peut-être approchée par son terme le plus grand (hypothèse "Winner-take-all"),
- les poids des lois gaussiennes sont égaux,
- les lois respectent une symétrie sphérique (i.e. $\Sigma_k = \sigma^2 I_p$ quel que soit k),

l'estimateur du maximum de vraisemblance $\Pi^* = \{\mu_1, \mu_2, ..., \mu_n\}$ est solution de l'équation des moindres carrés. Dans le cas où la distance euclidienne est utilisée, Π^* est similaire à un dictionnaire obtenu par Quantification Vectorielle (QV). En reprenant les hypothèses de S. Nowlan, nous avons utilisé l'algorithme LBG-splitting [6] pour estimer les dictionnaires, que nous identifions alors aux MSV. D'un point de vue théorique, ces modèles sont représentatifs du système vocalique phonétique que du système vocalique phonologique discret (voyelles prononcées en contexte et fortement coarticulées).

Pour l'identification, la phrase testée subit le même traitement que les phrases d'apprentissage (cf. paragraphe 3.1). La distorsion entre cet ensemble de voyelles $Y = \{y_1, y_2, ..., y_N\}$ et chacun des N_L MSV est calculée. Le langage le plus vraisemblable est celui qui minimise la distorsion.

4. EXPERIENCES

Les expériences ont été menées avec un ensemble de 5 langues issues du corpus de parole téléphonique multilingue d'OGI [5]. Ces langues (coréen, espagnol, français, japonais et vietnamien) ont été choisies par rapport à leurs systèmes vocaliques phonologiques [10]. Le japonais et l'espagnol ont des systèmes plutôt simples (5 voyelles) et très proches alors que les SV coréen et français sont plus complexes. Le système vocalique du vietnamien est quant à lui de complexité intermédiaire.

Les données d'OGI sont divisées en trois ensembles (apprentissage, développement et test). Les différents ensembles sont disjoints (il n'y a aucun locuteur en commun). Les ensembles de test et de développement comprennent approximativement 20 locuteurs alors que le corpus d'apprentissage en comprend 50. Les expériences préliminaires ont montré qu'il est souhaitable d'apprendre des modèles distincts pour les hommes et les femmes. Les expériences présentées ici ont été menées uniquement avec les locuteurs mâles à cause du faible nombre de locuteurs femmes (moins de 20 %) dans le corpus.

Plusieurs ensembles de paramètres ont été testés. Le système de référence utilise les 8 MFCC. La durée du segment et les dérivées des MFCC sont éventuellement intégrées. En effet, même si les voyelles sont des sons relativement stables il est connu qu'une modélisation dynamique est plus précise pour la parole spontanée. Lorsque les dérivées sont utilisées, un MSV dynamique est appris, et la distorsion prise en compte durant la reconnaissance est la distorsion conjointe des modèles statiques et dynamiques. En résumé, nous avons utilisé 4 modèles homogènes :

- #1 : MSV – 8 MFCCs
- #2 : MSVD – 8 MFCCs + durée
- #3 : ΔMSV – 8 ΔMFCCs
- #4 : ΔMSVD – 8 ΔMFCCs + durée

et 3 modèles composites :

- #5 : MSV + ΔMSV
- #6 : MSVD+ ΔMSV
- #7 : MSVD+ ΔMSVD

La taille des dictionnaires obtenus par QV a été arbitrairement fixée à 20 classes dans un premier temps puis elle a été étendue à 70 classes (modèles LBG-20 resp. LBG-70).

Les résultats (tableau 1) montrent que la prise en compte de la durée améliore les résultats des MSV statiques. Les ΔMSV améliorent aussi les scores d'identification lorsqu'ils sont utilisés avec les modèles statiques, même s'ils sont intrinsèquement moins discriminants. La différence entre les modèles #6 et #7, qui tient à la pondération de la durée, est moins significative. Ces ensembles de paramètres donnent les meilleurs résultats, avec un taux d'identification correcte de 62,7 % pour les phrases de 45 s. et 78,7 % pour 2 minutes de parole. Il est important de noter que l'information utilisée (les segments vocaliques) représentent en moyenne moins d'un tiers de la durée totale de la phrase.

Tableau 1: Taux d'identification correct – en fonction de la durée des phrases, des paramètres et des dictionnaires.

	LBG-20		LBG-70	
	45 s.	2 min.	45 s.	2 min.
#1	50.7	56.2	52.0	66.2
#2	57.3	73.7	56.0	76.2
#3	30.7	43.7	48.0	52.5
#4	36.0	43.7	46.7	58.7
#5	48.0	63.7	56.0	70.0
#6	58.7	**76.2**	58.7	**78.7**
#7	**61.3**	75.0	**62.7**	77.5

5. CONCLUSION

Les travaux présentés montrent que l'utilisation des systèmes vocaliques (SV) en IAL est efficace. L'extraction et la modélisation automatiques des SV ne requièrent aucun étiquetage manuel du signal, et les modèles obtenus par QV donnent des taux d'identification pour 5 langues satisfaisants (62, 7 %) pour une durée effective de signal traité de 15 secondes. Cette étude offre de nombreuses perspectives, en particulier quant à la modélisation spécifique d'autres catégories acoustiques (fricatives, plosives) moins sensibles aux variations inter-locuteurs. Enfin, l'exploitation des chemins symboliques issus des modèles acoustico-phonétiques dans un modèle de langage est aussi envisagée.

6. REFERENCES

[1] R. André-Obrecht, "A New Statistical Approach for Automatic Speech Segmentation", *IEEE Trans. on ASSP*, January 88, vol. 36, n° 1, pp. 29-40

[2] J. Hieronymus and S. Kadambe, "Robust spoken language identification using large vocabulary speech recognition", *Proc. ICASSP'97*, Munich, vol. 2, pp. 1111-1114

[3] M. Jardino, "Multilingual stochastic N-Gram Class Language Models", *Proc. ICASSP'96*, Atlanta

[4] L. F. Lamel and al., "Spoken Language Processing in a Multilingual Context", ICSLP'96, Philadelphia, pp. 2203-2206

[5] T. L. Lander et al., "The OGI 22 language telephone speech corpus", *Proc. Eurospeech '95*, Madrid, pp. 817-820

[6] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer", *IEEE Trans. On COM.*, January 1980, vol. 28 pp. 84-95

[7] I. Maddieson, *Patterns of Sounds*, 2nd Edition, Cambridge University Press, Cambridge 1984

[8] S. Nowlan, *Soft Competitive Adaptation: Neural Network Learning Algorithm based on fitting Statistical Mixtures*, PhD Thesis, School of Computer Science, Carnegie Mellon Univ. 1991

[9] F. Pellegrino and R. André-Obrecht, "From Vocalic Detection to Automatic emergence of Vowel Systems", *Proc. ICASSP '97*, München, pp. 1651-1654

[10] N. Vallée, *Systèmes vocaliques : de la typologie à la prédiction*, Thèse de 3ème cycle, Univ. Stendhal, 1994, Grenoble

[11] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", *Proc. IEEE Trans. On SAP*, January 1996, vol. 4, no. 1, pp. 31-44

Table ronde : conférence d'introduction

L'Association Européenne pour les Ressources Linguistiques (ELRA)

Khalid Choukri & Valérie Mapelli

ELRA - ELDA
55-57 rue Brillat Savarin 75013 Paris
Tél.: ++33 1 43 13 33 33 - Fax: ++33 1 43 13 33 30
e-mail: elra@calva.net - http://www.icp.grenet.fr/ELRA/fr/home.html

Abstract

The European Language Resources Association (ELRA) was established with the goal of promoting the creation, validation and distribution of language resources in Europe and to contribute to the development of the emerging market of language engineering. ELRA focuses on the identification and distribution of LRs and acts as an information centre using different media (quarterly newsletters, catalogue of language resources, Web site, announcements on various electronic lists). ELRA also managed to simplify the relationship between providers and users (customers) by drafting generic agreements specifying responsibilities and duties of each party. As of end December 1997, the ELRA catalogue listed about 66 speech resources, 40 monolingual lexicons, 70 multilingual lexicons, 16 written corpora and more than 360 terminological databases. Many resources still need to be identified. Moreover, and in order to add value to the resources it distributes, ELRA initiated the production of validation manuals. Partnerships with other organisations such as AUPELF and LDC, as well as several European funded projects (Parole, EuroWordNet, SpeechDat) will allow us to better serve the language engineering community.

1. Introduction

L'Association Européenne pour les Ressources Linguistiques (European Language Resources Association, ELRA) a été créée à Luxembourg en février 1995 afin de promouvoir les ressources linguistiques. La tâche première a consisté à répertorier les ressources linguistiques existantes et à informer les utilisateurs potentiels au niveau européen et international de leur disponibilité. La mission d'ELRA passe aussi par l'encouragement des utilisateurs à réutiliser des ressources existantes qui répondraient à leurs besoins avant d'entreprendre toute production, toujours coûteuse en temps et en argent.

Lors de sa première année d'activité, ELRA a accompli les tâches les plus urgentes pour le domaine de l'ingénierie de la langue. En plus des activités associatives, nous nous sommes concentrés sur l'identification et la négociation de ressources, la définition d'un modèle générique pour la rédaction de contrats de distribution et d'utilisation, la mise en place de moyens d'information (lettre d'information trimestrielle, catalogue de ressources linguistiques, site Web, annonces sur de nombreuses listes électroniques du domaine de l'ingénierie linguistique), ainsi que le lancement d'activités techniques pour la rédaction de manuels de validation de ressources.

2. Catalogue d'ELRA

Grâce au travail de l'agence de distribution ELDA, au concours du Bureau de l'association et à la contribution du panel d'experts responsable de l'identification des ressources, un premier inventaire a été réalisé et figure dans le catalogue d'ELRA, diffusé avec la lettre d'information trimestrielle et à travers le Web. Ce catalogue est organisé par catégorie de ressources. On y trouve la liste des corpus oraux, des lexiques monolingues, des lexiques multilingues, des corpus écrits, des bases terminologiques et un ensemble d'outils linguistiques génériques. Ce catalogue comporte uniquement les ressources pour lesquelles les propriétaires ont concédé un droit de distribution à ELRA. La première version, diffusée en mars 1996, comportait 22 bases orales, 6 bases écrites (corpus et lexiques) et 21 bases terminologiques.

Le dernier recensement (31 mars 1998) fait apparaître un total de 70 bases orales, 16 corpus monolingues et multilingues, 46 lexiques monolingues et 71 lexiques multilingues, plus de 360 bases terminologiques, la plate-forme logicielle et linguistique ALEP ainsi que les grammaires développées dans le cadre du projet LSGRAM.

3. Contrats avec les fournisseurs

La quête de nouvelles ressources a permis à ELRA de signer 59 contrats avec des fournisseurs (31 mars 1998). Un contrat peut rendre disponible via ELRA une ou plusieurs ressources. Le nombre de contrats relatifs à chaque type de ressource est indiqué dans le tableau ci-dessous.

Table 1 : Nombre de contrats par type de ressources (jusqu'au 31 mars 1998)

Ressources	Oral	Ecrit	Terminologie	Outils
Total	20	29	8	2

Dans un premier temps, on peut constater qu'en dépit du grand nombre de ressources disponibles, le nombre de contrats, et par conséquent, de fournisseurs reste très bas.

La répartition des fournisseurs par rapport aux différents pays de provenance est illustrée par le tableau suivant :

Table 2 : Nombre de contrats par pays (31 mars 1998)

Pays	Contrats
Allemagne	12
Royaume-Uni	7
France	6
Espagne	5
Italie	5
Pays-Bas	4
Belgique	3
Suisse	3
Danemark	2
Luxembourg	2
Grèce	1
Total	**50**

Le rôle que joue ELRA en tant qu'intermédiaire entre fournisseurs et utilisateurs permet d'éviter une multiplication de contrats. Prenons l'exemple du corpus multilingue MLCC. Celui-ci regroupe des données provenant de 6 journaux différents (et 6 langues différentes). ELRA a signé 1 contrat avec chaque fournisseur et permet à l'utilisateur de les obtenir en ne signant qu'un seul contrat d'utilisation avec ELRA au lieu d'un contrat avec chaque fournisseur, i.e. 6 contrats au total (et 6 systèmes juridiques différents) !

Les contrats de distribution types sont disponibles sur le site Web d'ELRA ou peuvent être obtenus en contactant les bureaux d'ELDA.

4. Distribution et commercialisation des ressources

Dans la mesure où ELRA joue un rôle d'intermédiaire entre le producteur/fournisseur et l'utilisateur, elle doit tenir compte des exigences et restrictions imposées par le producteur, en particulier en terme de compensation financière. De la même manière, la connaissance du marché et les contacts avec les acheteurs potentiels permettent à ELRA de remonter l'information concernant leurs demandes et leurs besoins. L'approche d'ELRA est donc de simplifier l'établissement des prix, de clarifier la définition des utilisations possibles et de réduire les restrictions qu'impose le producteur.

Le rôle d'intermédiaire suppose un paiement pour l'achat de chaque ressource, couvrant ainsi la compensation due au propriétaire. En effet, ELRA n'étant pas propriétaire des ressources, elle ne peut pratiquer une politique tarifaire qu'en coordination avec le propriétaire. La coordination avec ce dernier passe souvent par la mise à sa disposition de méthodes pour l'établissement de prix et l'étude des coûts de production, par la définition conjointe de la stratégie qu'il faut imaginer pour rentabiliser une ressource et de la cible d'acheteurs potentiels. Cette démarche permet de fixer un prix public ou une redevance. Dans le premier cas, le propriétaire et ELRA négocient la répartition des revenus. Dans le second cas, ELRA définit le prix de vente en rajoutant sa marge. La politique de distribution d'ELRA est basée sur sa volonté d'offrir des réductions importantes sur les prix de chaque ressource à ses membres. La marge est calculée pour que l'adhésion à l'association soit justifiée par les économies réalisées lors de l'achat de ressources.

Dans certains cas, les fournisseurs, en général des consortiums ayant reçu un support financier de la Commission Européenne ou des gouvernements nationaux, acceptent que leurs ressources soient distribuées gratuitement. Une bonne partie des frais de transport est alors répercutée sur le prix appliqué aux non membres. Une première analyse du catalogue des prix montre que nos membres bénéficient de réductions allant de 10% à plus de 70% sur le prix public. Dans des cas exceptionnels nous pourrions accepter de distribuer des ressources sans réduction consentie par le fournisseur à nos membres. Il s'agira là d'un service à rendre à nos membres, l'objectif étant qu'ELRA représente le site unique de service et de distribution des ressources.

Les restrictions imposées par les fournisseurs sont de deux types : une restriction relative aux utilisateurs et/ou une restriction relative à l'utilisation. Ainsi, des fournisseurs peuvent limiter la distribution à nos membres, d'autres aux organisations européennes, d'autres encore aux organismes de recherche, ou plus encore la limiter aux organismes de recherche "publics".

Quant aux restrictions liées à l'utilisation, il s'agit essentiellement de fournisseurs qui souhaitent offrir leurs ressources à des centres de recherche et refusent leur utilisation dans un but commercial.

5. Distribution et ventes

Les tableaux suivants récapitulent l'état de la distribution de ressources linguistiques par l'intermédiaire d'ELRA. La distribution aux membres et aux non membres, les ventes selon les différents collèges (oral, écrit, terminologie et outils) et les ressources distribuées à usage commercial ou de recherche sont représentées ci-dessous :

Table 3 : Ressources vendues au 22 janvier 1998

Membres	Non membres	Usage de recherche	Usage commercial
71	**19**	**39**	**51**

Table 4 : Statistiques sur les ventes aux membres et aux non membres (jusqu'au 22 janvier 1998).

Distribution	Ressources distribuées	Ventes réelles	Distribution gratuite
Membres	74	71	3
Non membres	28	19	9
Total	**102**	**90**	**12**

Les tables 3 et 4 soulignent la politique tarifaire d'ELRA à l'avantage de ses membres. En effet, le nombre de ressources distribuées aux membres triple par rapport à celui des ressources acquises par les non membres. A noter également l'intérêt des organisations pour l'usage des ressources à but commercial.

Table 5 : Statistiques sur les ventes selon les différents types de ressources (jusqu'au 22 janvier 1998).

Type de ressources	Ressources distribuées	Ventes réelles	Distribution gratuite
Oral	69	59	10
Ecrit	30	30	0
Terminologie	1	1	0
Outils	2	0	2
Total	**102**	**90**	**12**

Le tableau ci-dessus montre que les ressources terminologiques et lexicales n'ont pas fait l'objet de beaucoup de transactions, bien que de nombreuses demandes d'information nous soient parvenues et que plusieurs négociations soient en cours. Cela montre que la démarche des acquéreurs est différente entre les 3 collèges d'ELRA. En effet, vu les prix et les spécifications très pointues des ressources demandées, les acteurs impliqués dans les domaines de l'écrit et de la terminologie ne se décident pas facilement. Pour y remédier, ELRA révise actuellement les accords conclus avec ses fournisseurs afin d'offrir des prestations sur mesure en réponse à un cahier des charges rédigé par le demandeur.

Il va de soi qu'ELRA cherchera toujours à défendre l'intérêt de ses membres. Cela s'énonce clairement comme l'opportunité de disposer des ressources les plus adéquates au prix le plus faible possible et sans limitation quant à l'utilisation qu'ils peuvent en faire, sans pour autant violer les droits des fournisseurs.

Comme tout autre profit, les marges bénéficiaires issues de cette activité seront réinvesties dans les tâches de l'Association et permettront la création des ressources manquantes que les pouvoirs économiques classiques ne peuvent combler dans un délai raisonnable.

6. Adhésion à ELRA

L'adhésion à ELRA en tant que membre avec droit de vote, est réservée aux organismes, publics ou privés, légalement domiciliés en Europe. Dans un esprit d'ouverture, en particulier pour la mise à disposition de son catalogue de ressources linguistiques, aux organisations de recherche et industrielles en dehors de son champ d'action habituel, le Bureau d'ELRA, lors de sa réunion du 30 avril 1997, a décidé d'étendre les possibilités de devenir membre d'ELRA, sans droit de vote, aux organisations non européennes.

Lors de l'Assemblée Générale du 28 novembre 1997, les membres s'accordèrent pour modifier le montant de la cotisation annuelle comme suit : 750 ECU pour toutes les organisations à but non lucratif, 1000 ECU pour les PME européennes de moins de 50 employés, 1500 ECU pour les organisations européennes à but lucratif de plus de 50 employés, 5000 ECU pour les organisations non européennes à but lucratif.

Chaque membre choisit d'appartenir à l'un des trois collèges (écrit, oral ou terminologie) en fonction de son intérêt principal.

Les membres d'ELRA bénéficient de remises significatives sur les prix du catalogue de ressources (jusqu'à 70% de réduction). Ils peuvent obtenir des prix avantageux sur un certain nombre de publications techniques, de rapports commerciaux, etc. La lettre d'information trimestrielle d'ELRA est aussi diffusée à tous ses membres. D'autres publications, telles que des rapports d'études de marché, des documents traitant des aspects juridiques et commerciaux pour la distribution de ressources linguistiques sont également envoyés aux membres. Le site Web d'ELRA présente une section réservée aux membres où ceux-ci peuvent trouver toutes les informations nécessaires sur l'Association, ainsi que les événements récents traitant des projets européens et internationaux du domaine. Dans cette même perspective d'information, un bulletin électronique a été récemment créé afin de transmettre chaque mois les dernières actualités d'ELRA et des annonces pouvant intéresser les membres d'ELRA dans le cadre de leurs travaux.

7. ELRA et la francophonie

ELRA a réalisé en 1996, avec le support d'agences françaises chargées de la promotion des industries de la langue, une première analyse détaillée des corpus textuels et oraux dont une des composantes est le français. Pour ce faire, une analyse de notre catalogue a permis de dresser un état des lieux pour le français (étendu ensuite à chacune des langues de l'union européenne).

Ce travail s'est articulé autour de plusieurs tâches dont l'objectif était de rendre disponible le maximum de ces ressources ou de justifier l'impossibilité de le faire. Pour cela, nous avons procédé à l'identification de ressources auprès des acteurs travaillant directement ou indirectement sur la langue française, aussi bien lorsqu'il s'agit d'études monolingues que multilingues. Ensuite, nous avons cherché à classer cet inventaire selon plusieurs critères liés aux problèmes de disponibilité : disponibilité juridique (la possibilité d'obtenir des licences de distribution, de pouvoir céder ou concéder des droits d'utilisation ou d'exploitation pour des projets de recherche ou pour des développements de technologie), disponibilité commerciale (liée au désir ou à la stratégie du propriétaire de mettre une telle ressource sur le marché), disponibilité technique (un format d'encodage et de stockage standardisé ou suffisamment répandu est nécessaire pour que les échanges et les transferts se fassent de façon simple), l'intérêt technique et scientifique (l'adéquation de celle-ci avec une demande de la part des autres institutions, cette adéquation découle essentiellement de la pertinence de la base de données en terme de contenu même si un faible coût peut justifier l'acquisition d'une ressource peu importante pour un usage interne ou pour des tests). Cette étude a été finalisée en 1997 et devra permettre d'émettre des recommandations quant aux ressources manquantes.

Pour ce qui est de la distribution, il est à noter que les meilleures ventes de ressources contenant du français ont été acquises dans le domaine de la parole. Les ressources orales du français distribuées jusqu'à présent ont été les suivantes : BREF-80, BREF-Polyglot, BDLEX 50000, BDSONS, BDBRUIT et FRESCO.

8. Validation de ressources

La validation des ressources linguistiques est une tâche capitale. Les résultats des travaux de validation visent à permettre un meilleur contrôle des ressources. ELRA a ainsi travaillé sur un manuel de validation de ressources orales, et tentera de persuader les producteurs de l'adopter, celui-ci devant permettre de rendre leurs produits plus compétitifs sur le marché des ressources linguistiques. Le terme "validation" est utilisé en référence à l'activité de vérification de l'adéquation des ressources avec le marché et les standards, ainsi que le contrôle de qualité du produit. Le manuel de validation des ressources linguistiques est réalisé en synergie avec d'autres projets lancés lors du 4ème Programme Cadre de Recherche de la Commission.

Le manuel pour l'oral sera basé sur les travaux de SPEX, dans le cadre du projet SpeechDat. Une première version a déjà été présentée lors du Workshop COCOSDA à Rhodes (EuroSpeech, 26-27 septembre 1997) et sera rendue disponible dès que le résultat des discussions aura été incorporé. Ce manuel permettra de définir un ensemble initial de critères de validation formels, ainsi que des directives méthodologiques générales. La validation technique comprendra notamment la taille du corpus, le nombre de locuteurs enregistrés, le type de parole (spontanée, lue, monologue, dialogue, groupe de discussion, etc.), les signaux enregistrés (audio, vidéo, etc.), la manière dont les signaux ont été enregistrés (analogique ou digital), l'annotation venant avec les signaux, le support technique sur lequel les données sont fournies, les jeux de caractères utilisés, et la documentation annexe.

9. Accords de coopération

Afin de mieux répondre aux attentes de ses membres et de ses sponsors, ELRA noue des accords de partenariat avec d'autres organismes dont la mission est complémentaire à la sienne. On peut citer à titre d'exemple ELSNET (European Language and Speech Network), dont nous avons repris la tâche de distribution de ressources linguistiques, l'association PAROLE qui nous confiera, pour les distribuer, les ressources écrites actuellement en développement, TELRI (projet Copernicus pour la production des ressources écrites en Europe de l'Est), BABEL (projet Copernicus pour la production des ressources orales en Europe de l'Est), l'AUPELF (distribution des ressources écrites et orales pour les projets francophones d'évaluation), LDC (Linguistic Data Consortium) aux Etats-Unis et EDR au Japon.

10. Perspectives 98

Dès sa fondation, ELRA a concentré tous ses efforts sur la mise en place d'une infrastructure opérationnelle apte à mettre en oeuvre sa politique de distribution. Ensuite, ELRA s'est consacrée à la rédaction des contrats génériques qui permettent d'avoir une base de négociation avec les fournisseurs potentiels. Grâce à cela, un premier ensemble de ressources linguistiques a été identifié et sa distribution assurée. Cette année, ELRA prend une nouvelle orientation qui se décline selon quatre thèmes : l'organisation de la première conférence internationale sur les ressources linguistiques et l'évaluation, la mise au point d'autres manuels de validation, l'exploitation des résultats de l'étude de marché réalisée en 1997 et l'intégration, dans sa réflexion, de la distribution des ressources par réseau.

ELRA organise du 28 au 30 mai 1998 la première conférence internationale sur les ressources linguistiques et l'évaluation. La conférence sera accueillie par l'Université de Grenade avec le concours de la DG XIII de la Commission Européenne. L'annonce de cette conférence a suscité un grand intérêt puisque ELRA a reçu 250 propositions d'articles et 12 propositions d'ateliers.

La seconde tâche permettra d'offrir un label de qualité aux ressources que l'on distribue. Ainsi, les manuels nécessaires à la validation des ressources écrites (aussi bien des lexiques que des corpus), ont pu être rendus disponibles récemment grâce à des experts fortement impliqués dans ce domaine. Pour les ressources orales, une bonne partie du travail a été réalisée dans le cadre du projet SpeechDat(M). ELRA est actuellement en train de capitaliser sur les résultats de ce projet afin de mettre au point un manuel et des outils permettant la validation des bases de données téléphoniques dans un premier temps, et un manuel plus général dans un second temps.

La troisième tâche d'ELRA devrait permettre de mieux cerner les besoins et les attentes de nos membres et de nos clients suite à l'étude de marché réalisée en 1997. ELRA collecte déjà des informations sur le marché des ressources linguistiques à travers le monde en tirant parti de sa propre expérience et de celle de ses membres. Mais une étude spécialement dédiée à cette tâche a permis de consolider nos orientations.

Le quatrième thème correspond à une réflexion quant à l'utilisation des réseaux informatiques pour la distribution des ressources linguistiques. ELRA envisage de profiter de tous les outils en cours d'élaboration dans le cadre du "commerce électronique". Dans cette perspective une coopération avec SILFIDE est en cours de discussion.

11. Remerciements

Nous souhaitons profiter de cette publication pour remercier la DG XIII de la Commission Européenne, la Délégation Générale à la Langue Française (Ministère français de la Culture), le Ministère français de l'Enseignement Supérieur et de la Recherche, ainsi que le Ministère français de l'Industrie, pour leur soutien.

Conférence invitée

ANALYSE EN ONDELETTES ET RECONNAISSANCE DE LA PAROLE.

Stéphane H. Maes *IBM*, T. J. Watson Research Center,
Human Language Technologies, Acoustic Processing Department,
Room 23-148B, P.O. Box 218, Yorktown Heights, NY 10598, USA.
Phone: +1-(914)-945-2908
Fax: +1-(914)-945-4490
e-mail: smaes@watson.ibm.com

Journées d'études de la parole, *JEP*,
Martigny, Suisse,
Juin 1998

RESUME

L'analyse en ondelettes et ses extensions en analyses à multi-résolution ont mis moins de quinze ans pour s'imposer comme incontournables outils d'analyse mathématique et numérique du signal. Alors qu'en analyse numérique de signaux et en codage numérique, les ondelettes sont régulièrement utilisées au coeur de nombreux algorithmes, en reconnaissance de la parole, leur usage est encore très marginale.

Dans ce papier, nous présentons différentes repr'esentations du nerf auditif avantageusement mathématiquement modélisées par ondelettes.

Les modèles de production de la parole conduisent généralement à l'introduction de modèles de modulation qui représentent les signaux de paroles à l'aide de composants primaires qui sont des sinusoides modulées en amplitude et en phase.

Depuis leurs débuts comme transformations continues et le développement des représentations en *ridge-skeleton*, les ondelettes se sont avérrées bien adapatées à l'extraction des lois de modulation de composants uniques ou suffisamment distincts. Cependant, les approximations introduites par ces méthodes limites leurs possibles applications à de réels signaux de paroles. Les similarités entre la transformation en ondelette de composant primaires et les représentations de la parole dérivées de modèles du nerf auditif tels que *EIH* conduisent à l'introduction de mesures synchrosqueezées.

Ce papier présente une nouvelle transformation du plan temps-fréquence obtenu par analyse quasi-continue en ondelettes. La représentation synchrosqueezée n'est pas limitée a l'analyse de la parole. De plus, cette représentation est essentiallement indépendante de l'ondelette génératrice d'analyse.

Notre analyse utilise la phase du signal. Elle décrit naturellement des modèles du nerf auditif comme *IFD*, *SBS* et *EIH*.

Basée sur une robuste représentation sur-complète du signal, l'analyse peut être inversée afin d'extraire les composants primaires d'un signal de parole. **Mots clefs:** transformation en ondelettes, analyse temps-fréquence, représentation du nerf auditif, analyse numérique du signal, reconnaisance automatique de la parole, reconnaisance automatique du locuteur, modèle de modulation.

1. INTRODUCTION

En général, les paramètres at méthodes rencontrés en codage et reconnaissance de la parole sont obtenus par estimation de densitées spectrales. C'est le cas de l'analyse spectrale, avec ses variations en échelle de Mel or de Bark, le cepstre, obtenu par *FFT*, *LPC* et dérivatifs, la function d'autocorrélation, le contenu en énergie et les dérivée temporelles (coeffificent deltas et delta-delta) [24, 23, 25, 67]. C'est aussi le cas du modèle de modulation par formants [39, 14, 68, 4] et de la représentation en sinusoides [61].

Très peu de modèles utilisant la phase ont été proposés dans la littérature Dans le cadre de ce papier, les principaux modèles sont: *EIH* (Ensemble Interval Histogram) [27, 28, 67], *SBS* (in-Synchrony Bands Spectrum) [26] et *IFD* (Instantaneous-Frequency Distribution) [22]. Ces trois modèles sont dérivés de modèles temporels mais non locals du nerf auditif [27, 28]. D'autres modéles ont été proposés pour caractériser avec des ondelettes certains comportements du nerf auditif [8].

Nous commencons par le modèle classique de production de la parole avec sources et filtres. Cela nous permet de justifier l'introduction du modèle de modulation. Un rapide survol de la transformation en ondelettes et de ses applications à l'extraction de composant primaires, illustre les limitations d'un usage direct de representations temps-échelles. En introduisant l'approche *EIH*, nous découvrons un différent point de vue. Transposé à l'analyse en ondelettes, ces nouveaux principes débouchent sur les concepts de plan sychrosqueezé.

Il y a différentes raisons pour lesquelles nous avons décidé d'utiliser la transformation en ondelettes. Utilisée avec une ondelettte génératice complexe, la transformation en ondelettes contient d'intéressantes informations dans sa norme et sa phase [18, 17, 12, 19, 35, 58]. L'information contenue dans la norme est semblable au contenu de la densité spectrale. Par contre la phase est bien plus indépendante du l'amplitude du signal. De plus, le premier étage du système auditif humain est simulé par un banc de filtres à facteur de qualité constant [1] [27, 67]. La transformation en ondelette est aussi construite avec un banc de filtres à facteur de qualité constant, à un facteur de normalisation près. Finalement, le facteur de dilatation produit un support adaptatif en temps-fréquence. Un tel support est particulièrement adéquat pour étudier des phénomènes vivant à différentes échelles. C'est typiquement le cas des signaux de parole pour lesquels des spectrogrammes à courtes ou longues fenêtres temporelles illustrent différent aspects importants du signal. modifications du cepstre classique en remplacant l'analyse de Fourier par une analyse en ondelettes ou une représentation similaire.

2. LE MODELE DE MODULATION

Un modèle de sourcse et filtres légèrement différent de ce qui est généralement présenté dans la littérature est illustré en figure 1. Il s'agit d'un circuit linéair équivalent qui modélise les mécanismes de production de parole. Ses limitations sont discutées dans [21, 24, 16].

Pour les sons voisés, un train d'impulsions périodiques est filtré pour simuler l'effet de glotte. Cela produit la composante périodique de l'excitation du signal. La composante non-périodique est obtenue en ajoutant une source de bruit blanc, généralement absente dans la littérature. Pour des sons non-voisé, seule la source de bruit blanc est présente. Le signal d'excitation traverse le conduit vocal caractérisé par une fonction de transfert $H_{vt}(z)$ [2].

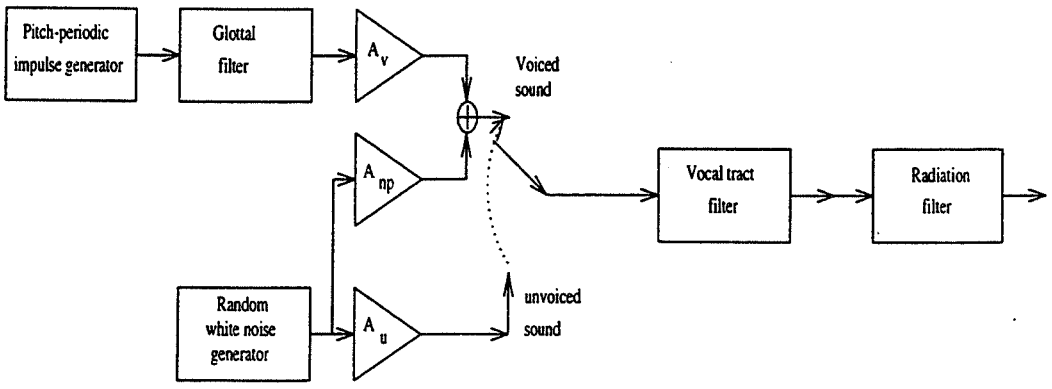


Figure 1: Modèle avec sources et filtres de producion de la parole. A_v est le gain de la partie voisée, A_u est le gain non-voisé et A_{np} la composante non-voisée dans le signal voisé.

[1] Le facteur de qualité d'un filtre est défini comme le quotient de sa fréquence centrale à sa bande passante.

[2] Rigoureusement, z devrait être la variable de Laplace puisque le signal de parole est continu. Cependant la relation entre transformation de Laplace et transformation en z est bien connue [11]. Cette notation est d'autant bien justifiée que le signal de parole est toujours echantillonné avant d'être traité.

Un terme supplémentaire $R(z)$ modélise l'effet de radiation de la bouche. $R(z)$ a toujours un effet de filtre passe-haut [10, 64, 21, 20, 62, 63].

Un tel system présente des contributions dominantes en pression ou vitesse aux fréquences de résonnances et anti-résonnances du conduit vocal. Ces contributions dominantes sont les modes propres de la fonction de transfert. Cette approche introduit naturellement le concept de modèle de modulation. Cette interprétation is particulièrement bien adaptée au sons voisés, pour lesquels les composants primaires, c'est 'a dire les modes propres, sont directement liés à la définition classique de formants. Cependant, les sons non-voisé satisfont également au modèle.

Le modèle de modulation considère qu'un signal de parole peut être écrit comme combinaison linéaire de composants modulés en amplitude et phase: les composant primaires;

$$f(t) = \sum_{k=1}^{N_{pc}} A_k(t)\, cos(\Phi_k(t)) + \eta(t) \tag{1}$$

où $A_k(t)$ est l'amplitude instantanée du k^{eme} composant primaire et $\Phi_k(t)$ est la phase instantannée. $\eta(t)$ prend en compte le bruit additif et les erreures du modèle, en particulier celles dues a la somme finie. Ce modèle a également ete proposé dans la littérature sous le nom de "elementary waveform speech model" [39, 14].

Pour des signaux de parole, les premiers composants primaires sont le signal d'excitation produit par la glotte (pitch) ainsi que ses premiers composants harmoniques. Ces composants sont généralement négligés par de classiques algorithmes qui amplifient sélectivement le signal [51, 4]. Par définition, ces composants ne sont pas considérés comme formants. Les autres composants primaires englobent naturellement les formants d'un signal de parole. Les formants caractérisent relativement bien un signal voisé. Cependant, à côté du pitch, harmoniques et formants, le modèle de modulation introduit souvent d'autres composants primaires qui résultent d'interférences entre les différents composants primaires et leurs intéractions au travers du conduit auditif. Suivant l'application, différents nombres de composants sont nécessaires. En reconnaissance de la parole ou du locuteur, les composants sont généralement limités aux formants classiques [4]. Par contre, en codage, les coders par analyse et synthèse basés sur la représentation sinusoidale utilisent davantage de composants primaires, même dans le cas de sons non-voisés [61, 36, 59, 60, 66, 2, 26]. La possibilité de représenter des sons non-voisés présente une justification supplémentaire pour notre choix de ne pas distinguer entre sons voisés et non-voisés. Pour des sons voisés, la plupart des composants disparaissent. Les composants restants présentent de larges bandes passantes instantannées et juxtaposent leurs voisins. La fréquence centrale devient difficile à extraire. L'extraction des composants primaires impose co-curvilinéarité avec les régions voisées voisines. Par définition d'un son non-voisé, les premiers composants (pitch et premiers harmoniques) sont absents ou fortement atténués. Cependant, le modèle de modulation demeure valide comme décrit dans [5, 14]. Essentiellement, il peut être décrit comme si l'excitation est différente pour chaque composant primaire. Cela peut être vérifié par les amplitudes instantanées comme expliqué dans [48, 52, 54].

Pour une signal donné, la tâche adressée par ce travail est de minimizer la norme Euclidienne dans $L^2(\mathbb{R}, dx)$ de $\eta(t)$. Comme pour le problème de représentation analytique, la solution à notre problème n'est pas unique.

3. TRANSFORMATION EN ONDELETTES

Il est utile de présenter rapidement les notations utilisées dans ce papier. WT, la transformation en ondelette d'un signal $f(t)$ utilisant $g(t)$ comme ondelette génératrice d'analyse est définie comme:

$$\mathcal{WT}_f(a,b,g) = < g_{(a,b)} | f > \tag{2}$$

où,

$$g_{(a,b)}(t) = \frac{1}{\sqrt{a}} g(\frac{t-b}{a}) \tag{3}$$

$a \in \mathbb{R}_+ \setminus \{0\}$ est le paramètre de dilatation et $b \in \mathbb{R}$ est le coefficient de translation temporelle[3]. a est souvent assimilé à l'invere de la fréquence tandis que b est considéré comme la variable temporelle. $< .|. >$ est le classique produit scalaire. La transformation en ondelette à une échelle donnée est la projection de $f(t)$ sur des translatée de $g_{(a,0)}(t)$[4].

[3]Dans la littérature, une autre convention est souvent utilisée où $a \in \mathbb{R} \setminus \{0\}$.

[4]La projection de $f(t)$ sur l'espace sous-tendu par l'ensemble $\{e_k(t)\}_k$ est definie comme: $\{P_{\{e_k(t)\}}\{f(t)\}\}_k \triangleq \{< e_k | f >\}_k$. Quand nous utilisons le terme de "projection" au lieu de projection orthogonale, le lecteur devrait toujours utiliser cette définition. L'absence de de normalisation par la norme de $e_k(t)$ permet de sélectionner différentes normalisations pour la transformation en ondelettes. Par exemple, certains auteurs utilisent cette normalisation: $g_{(a,b)}(t)$:

$$\mathcal{WT}_f(a,b,g) = \frac{1}{\sqrt{a}} < g_{(a,b)} | f > \tag{4}$$

Tout signal $f(t)$ dans un espace de Hilbert satisfaisant $\mathcal{H}$ peut être reconstruit à partir de sa transformée en ondelette WT, utilisant une ondelette génératrice d'analyse $h(t)$. Cette propriété est appelée *résolution de l'identité*,

$$\forall f(t) \in \mathcal{H} \; : \; f(t) \; = \; \frac{1}{C_{g,h}} \int_{-\infty}^{+\infty} \int_{0}^{+\infty} h_{(a,b)}(t) \; \frac{db \; da}{a^2} \; \mathcal{WT}_f(a,b,g) \tag{5}$$

Elle requière de respecter la *condition d'admissibilité* [5]:

$$0 \; < \; \left| C_{g,h} \; = \; \int_{0}^{+\infty} \frac{\hat{h}(\omega)\hat{g}^*(\omega) \; d\omega}{\omega} \; = \; \int_{0}^{+\infty} \frac{\hat{h}(-\omega)\hat{g}^*(-\omega) \; d\omega}{\omega} \right| \; < \; +\infty \tag{6}$$

Ces équations sont écrites dans le cas continu bi-orthogonal: l'ondelette génératrice d'analyse peut être choisie différente de l'ondelette génératrice de synthèse [37].

4. LA TRANSFORMATION EN ONDELETTE APPLIQUEE AU MODELE DE MODULATION

4.1 Techniques d'extraction de simples fonctions

Un des principaux avantages de la transformation en ondelettes est sa capabilit'e de détecter, extraire et manipuler une courbe. Dans cette section, l'application de la transformation en ondelette d'une simple courbe est rapidment décrite.

La fonction génératrice d'analyse, utilisée dans ce papier, est supposée analytique:

$$\omega \; \leq \; 0 \; \Rightarrow \; \hat{g}(\omega) \; = \; 0 \tag{7}$$

4.1.1 Etude 'une exponentielle complexe

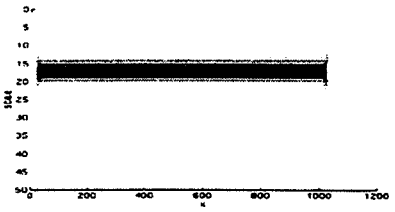


Figure 2: Module de la transformation en ondelette d'une exponentielle complexe. L'axe horizontal est échantillonné à 8 kHz. L'axe vertical représente l'indexe des différentes sous-bandes, arrangées comme expliqué dans [48, 44, 47]. Il y a 5 octaves, avec 7 sous-octaves , ou voix, également espacées et une octave supérieure interpolée. Les indexes inférieurs sont associés aux hautes fréquences. Un traitement aux limites a été appliqué, mais les contributions parasites aux bords n'ont pas ét'e éliminées [48, 46, 44].

Le signal est supposé être une exponentielle complexe isolée,

$$f(t) \; = \; A_f \; e^{j(\omega_f t \; + \; \Phi_f)} \tag{8}$$

Sa transformation en ondelette est

$$\mathcal{WT}_f(a,b,g) \; = \; \sqrt{a} A_f \hat{g}^*(a\omega_f) e^{j(\omega_f b \; + \; \Phi_f)} \tag{9}$$

Ecrivons la transformation en ondelette en repésentation phasorielle

$$\mathcal{WT}_f(a,b,g) \; \triangleq \; |\mathcal{WT}_f(a,b,g)| \; e^{j \; \Phi_{\mathcal{WT}_f}(a,b,g)} \tag{10}$$

Une pseudo-fréquence est introduite comme étant

$$\omega_{\mathcal{WT}_f}(a,b,g) \; = \; \frac{d\Phi_{\mathcal{WT}_f}}{db}(a,b,g) \tag{11}$$

[5] $\hat{g}(\omega)$ est la transformation de Fourier de $g(t)$. Elle est normalisée de la façon suivante: $\hat{g}(\omega) \; \triangleq \; \int_{-\infty}^{+\infty} g(t) \frac{e^{-j\omega t}}{\sqrt{2\pi}} \; dt$

Bien qu'associée à une ligne spectrale unique, la transformation en ondelette produit un ruban de largeur fréquentielle essentiellement de taille $[\frac{\omega_l}{\omega_f}, \frac{\omega_h}{\omega_f}]$ où $[\omega_l, \omega_h]$ désigne le support fréquentiel essentiel de $\hat{g}(\omega)$. La figure 2 presente le module de la transformation en ondelette du signal (8).

En d'autres mots, l'information que nous espérions trouver concentrée sur un ligne droite est répandue sur un ruban qui dépend de l'ondelette génératrice d'analyse. Cela illustre clairement notre affirmation que la transformation en ondelette n'a pas de directe signification physique. Cela justifie la recherche d'une représentation alternative qui ne dépendrait pas de l'ondelette g'enératrice d'analyse.

Une exponentielle complexe satisfait l'équation suivante [6]:

$$\omega_{WT_f}(a, b, g) = \omega_f \tag{13}$$

Des cas plus complexes sont étudiés dans [48, 53, 58, 34, 31, 35].

La figure 3 montre le module de la transformation en ondelette d'un signal de parole disant $/\underbrace{\cdots - a - a}_{} - \underbrace{i - i - \cdots}_{}/$. Chaque composant primaire de ce signal voisé présente le même comportement qu'une exponentielle complexe isolée.

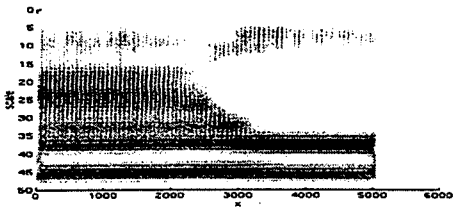


Figure 3: Module de la transformation en ondelette de $/\underbrace{\cdots - a - a}_{} - \underbrace{i - i - \cdots}_{}/$. Un bruit coloré est présent avec SNR = 15 dB. L'axe horizontal est échantillonné à 8 kHz. L'axe vertical représente l'indexe des différentes sous-bandes, arrangées comme expliqué dans [48, 44, 47]. Il y a 5 octaves, avec 7 sous-octaves , ou voix, également espacées et une octave supérieure interpolée. Les indexes inférieurs sont associés aux hautes fréquences. Un traitement aux limites a été appliqué mais les contributions parasites aux bords n'ont pas ét'e éliminées [48, 46, 44]. L'ondelette génératrice d'analyse est donnée par (17).

4.2 La méthode du ridge-skeleton

La méthode du *ridge-skeleton* est décrite en détails dans [18, 17, 19, 69, 29, 32, 33]. En appliquant les méthodes d'expansions asymptotiques d'intégrales oscillants rapidement, en paticulier la *steepest descent method*, nous découvrons que les points stationnaires t_s de la phase de l'intégrand produisent les contributions dominantes des intégrales associées à la transformation en ondelettes [13, 38]. Les *ridges* sonts les courbes associ'ees, dans le plan temps-échelle, aux points stationnaires de la transformation en ondelettes. Les *ridges* permettent de directement déduire les fréquences instantannées dans le plan temps-fréquence.

Le *skeleton* est défini comme la restriction aux *ridges* de la transformation en ondelettes du signal. Il peut être estimé par des expressions analytiques.

Pour des ondelettes génératrice d'analyses qui sont analytiques et asymptotiques et pour des ondelettes génératrice d'analyse qui sont d'ampliudes Gausiennes, le *ridges* et le *skeleton* sont indépendant de l'ondelette génératrice. Par contre, les erreurs d'approximations dépendent du signal et de l'ondelette génératrice d'analyse.

La connaissance du *skeleton* et du *ridge* permet l'extraction des amplitudes instantannées. Des algorithmes ont été proposés pour extrare le *ridges* sans devoir calculer la transformation en ondelettes sur tout le plan temps-fréquence.

Aussi longtemps que les différents composants primaires ne sont pas top proches les uns des autres, les contributions dominantes proviennent des régions du plan temps-fréquence où l'ondelette oscille à la même fréquence que le composant le plus proche.

Dès lors, pour toutes les valeurs de b pour lesquelles il n'y a pas de composant se superposant, pas de croisements ni de composants s'approchants trop près les uns des autres , N_{pc} *ridges* sont extraits. Aussi longtemps qu'il n'y a pas d'interférences

[6] L'équation (13) est toujours satisfaite au second ordre par une exponentielle à phase modulée. Elle devient:

$$\frac{d\Phi_{WT_f}}{db}(a, b, g) = \omega_f(b) = \frac{d\Phi_t}{db}(b) \tag{12}$$

Cela peut être justifié en utilisant la *steepest descent method* [13, 38]. Ceci justifie l'extension de notre méthode a n'importe quelle loi de modulation de phase.

les courbes qui résultent décrivent les lois de modulation de fréquence des différents composants. Lorsque certains composants s'approchent trop les uns des autres, au moins deux *ridges* se mélangent et en raison du phénomène bien connu de battements, la *steepest descent method* n'extrait plus les contributions dominantes de la transformation en ondelettes. Dès lors, les *ridges* non seulement se mélangent mais ils commencent à errer sans réel relation avec les lois de modulation des composants primaires associés.

Il est possible d'utiliser des méthodes perturbatives pour interdire le croisement des *ridges*. Elles résultent de ce que l'information décrivant un composant primaire est présente non seulement sur le *ridge*, juste une courbe dessinée dans le plan temps-fréquence, mais aussi sur le ruban qui s'étend autour du *ridge*. Dès lors, en dessinant deux courbes perturbées autour des *ridges*, nous extrayons suffisamment d'information pour estimer les lois de modulation de fr'equence et d'amplitude.

5. ENSEMBLE INTERVAL HISTOGRAM

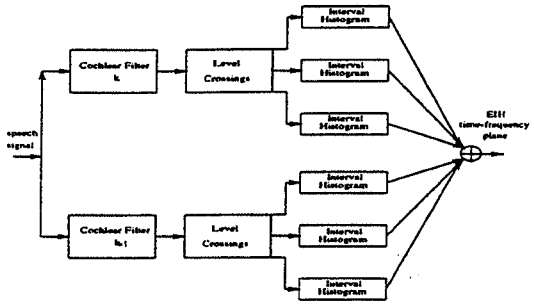


Figure 4: Diagramme décrivant le modèle *EIH*. Le banc de filtre modéise la réponse en fréquence des *inner hair cells*. Les deux blocs suivant implémentent l'effet du nerf auditif. A chaque instant, les histogrammes associés à chaque niveau dans chaque sous-bandes sont additionnés pour produire la repr'esentation *EIH*.

Des descriptions détaillés du système auditif périphérique humain peuvent être trouvées dans [67, 24, 21, 10, 28, 1, 30]. La représentation *EIH* est une tentative de tirer parti du déclenchement synchronisé des neurones qui contiennet l'information fournie aux étages supérieurs du nerf auditif [7]. En d'autres termes, il est concevable que pour analyser un signal de parole, du moins en reconnaissance de parole, il est suffisant de ne considérer que les caractéristiques d'un signal qui produisent les mêmes déclenchements.

Comme illustrés par la figure 4, la réponse en fréquence de la membrane basilière est modélisée par un banc de filtre cochléen. Le filtre est esentiellement à facteur de qualité constant si ce n'est pour les premiers 1000 Hz [67, 27, 28]. Les étages suivants modélisent les non-linéarités du déclenchement des neurones le long du nerf auditif: les instances de level-crossings simulent la transduction de mouvement en activité neurale au niveau des inner hair cells. Le modèle est considéré non-localisé/temporel puisqu'il assume que aucune information n'est transmise à propos de la location des inner hair cells le long de la membrane basilière. Seul l'instant relatif de déclenchement es enregistré. C'est ce que l'histogramme d'intervales modélise. L'addition mesure l'étendue de la cohérence de l'activité neurale le long du nerf auditif.

L'algorithme peut être réinterpr'eté de la façon suivante. Le signal est décomposé en sous-bandes à l'aide d'un banc de filtre d'ondelettes hybrides [8]. Dans chaque sous-bande, les fréquence dominantes sont extraites. A chaque instant t et pour chaque fréquence extraite, une mesure accumule les contributions de chaque sous-bandes dans lesquelles cette fréquence est d'etectée au temps t.

6. LA REPRESENTATION DU PLAN SYNCHROSQUEEZEE

Comme déjà expliqué, la transformation en ondelettes produit une représentation du signal qui dépend de l'ondelette génératrice d'analyse. Cependant, il est possible d'extraire de cette représentation des paramètres qui ne dépendent essentiellement que du signal. Les méthodes existantes dessinent des courbes de contributions dominantes dans le plan temps-fréquence d'où elles produisent la seule information utilisée pour caractériser le signal. Les paramètres extraits ne dépendent pas explicitement de l'ondelette génératrice d'analyse.

[7] Des mod'eles similaires ot ét'e proposés préalablement: instantaneous-frequency distribution (*IFD*) [22] et in-synchrony bands spectrum (*SBS*) [26].

[8] Dans notre contexte, la terminologie banc de filtre en ondelettes hybrides signifie que le filtre générateur peut changer de niveau en niveau dans une implémentation hiérarchique [45].

La leçon la plus importante est probablement que les lois de modulation de phase sont totalement contenues dans la phase de la transformation en ondelettes. Cette information peut être facilement sépar'ee de la l'influence de l'amplitude qui décrit davantage les lois de modulation d'amplitude.

A ce stade, il devrait être clair que l'information à propos d'un composant primaire est distribuée sur le ruban qui lui est associé. C'est une directe conséquence de l'équation (9). Dès lors, au lieu de sélectionner des courbes, il est concevable de concentrer le ruban sur une courbe. Cela évite toute perte d'information. De plus, la représentation ne dépend plus de 'ondelette génératrice d'analyse. Finalement, la représentation est plus robuste.

Ce concept de regroupement de l'information contenue dans des sous-bandes qui ont le même comportement est semblable à l'approche du modèle *EIH*. Il nous reste à définir l'extraction de fréquences et la mesure d'accumulation d'information.

La nouvelle représentation proposée dans ce papier est applicable à la transformation en ondelettes, à la transformation de Gabor ou de facco générale à toute distribution temps fréquence. qui contient suffisament d'information à propos des lois de modulation de fr'equence du signal. Des généralisations sont pr'esent'ees dans [48, 46, 52, 53].

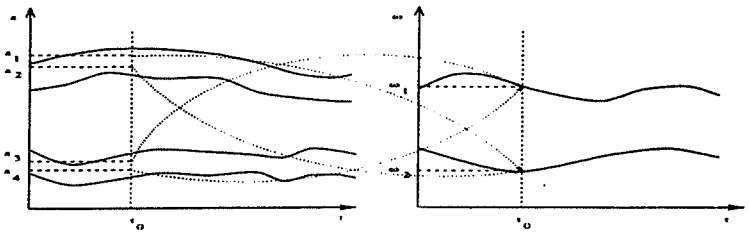


Figure 5: Si chaque ruban est associé à seulement une pseudo-fréquence, les rubans concentrés dégénèrent en deux courbes.

La mesure synchroqueez'ee associée avec la transformation continue en ondelettes est introduite [48, 53, 15] [9],

$$\mu_{Synmod}(\omega,b,f,g) = \left| \int_{(b,\omega_{WT_f}(a,b,g)) \in \mho(b,\delta_b,\omega,\delta_\omega)}^{SIS} WT_f(a,b,g)\, w_p(a)\, da \right| + \int_{(b,\omega_{WT_f}(a,b,g)) \in \mho(b,\delta_b,\omega,\delta_\omega)}^{SIS} |WT_f(a,b,g)|\, w_m(a)\, da \quad (14)$$

où

$$\mho(b,\delta_b,\omega,\delta_\omega) = [b - \delta_b(\omega), b + \delta_b(\omega)] \otimes [\omega - \delta_\omega, \omega + \delta_\omega] \quad (15)$$

δ_ω est la moitié de la résolution en fréquence qui peut être obtenue par cette approche et $\delta_b(\omega)$ est la moitié de a longueur de l'intevalle utilisé pour adoucir la mesure dans la direction b [10]. C'est lié à la réallocation temporelle décrite dans [6, 7]. $w_p(a)$ et $w_m(a)$ sont des facteur de pondération.

Dans le cas d'une (Quasi-Continuous Wavelet Transform) ($QCWT$) [11]. Les distributions sont implément'ees avec des sommes finies [53, 48, 41, 57, 56, 43, 42, 47, 15].

---

[9] Les limites de l'intégrale sont symboliques puisque qu'une fréquence isolée est de mesure nulle. Cette notation doit toujours être comprise comme étant donnée par un ensemble de mesure non-nulle paramétrisé par ω.

[10] Le choix le plus évident est d'utiliser $\delta_b(\omega) \propto \frac{1}{\omega}$.

[11] La $QCWT$ est une transformation à temps discret le long de l'axe temporel sans sous-échanillonage temporel et avec la possibilité de choisir n'importe quel grille d'échantillonage le long de l'axe des fréquences. En fait, un échantillonage trés utile pour l'analyse est défini par

$$a = a_0^{-m - \frac{\epsilon}{n_{so}+1}}, \quad a_0 > 0, a_0 \neq 1, \quad m \in \mathbb{Z}, \quad \epsilon = 0, \cdots, n_{so} \quad (16)$$

où $n_{so} \in \mathbb{N} \setminus \{0\}$.

Dans le cas de $QCWT$, a_0 est généralement choisi égual à 2. Dans ce cas particulier, les différent niveaux, indexés par le paramètre m sont arrangés en octaves et les niveaux intermédiaires sont arrangés en sous-octaves. Pour d'autres choix de a_0, la ême terminologie est utilisée. Cependant ces choix ne présentent que peu d'intéret, puisque la grille d'échantillonage (16) devait toujours offrir suffisamment de résolution en 'echelle.

Si $a_0 \neq 2$, les ondelettes génératrices d'analyse ne sont généralement pas des versions dilatées d'une même ondelette génératrice d'analyse. Cependant, ce n'est pas une issue comme nous développons une représentation essentiellement indépendante de l'ondelette génératrice d'analyse.

Dans ce papier, toutes les expériences ont t'e réalisés avec $a_0 = 2$. Des algorithmes rapides de $QCWT$ pour analyse et synthèse sont décrits dans [49, 55, 48].

Comme une exponentielle complexe a toujours sa contribution étendue sur le support fréquentiel essentiel de l'ondelette dilatée en conséquence, seul un interval connexe principal en a doit être pris en considération das l'intégrale de l'équation (14). Toutes les autres contibutions doivent être rejetées. Cette technique est appel'ee selective increment strategy (*SIS*). Ele réduit la perturbation du plan synchrosqueezé introduite par du bruit et dès lors augmente la résistance au bruit et la stabilité de n'importe quelle méthode qui l'utilise. La notation $\int^{SIS}$ décrit la compatibilité de la mesure associée avec *SIS*: l'intégrale est soumise à *SIS*.

Nous utilison des valeurs complexes au niveaux des mesures de telle façon que lorsque deux composants primaires se rapprochent de trop, ils présentent diffrents comportements de phases dans leurs transformations en ondelettes. Dès lors, leurs interférences sont incohérentes et une claire séparation appara^entre les deux composants.

Cette formulation est également très utile pour avoir une reconstruction sélective basée sur l'algorithme de fusion sélective décrit plus tard et la stratégie de sommation directe décrite dans [48, 49].

7. PLAN SYNCHROSQUEEZE DE SIGNAUX DE PAROLE

Les figures 6 à 8 présentent le plan squeezé et synchrosqueezé obtenu sans *SIS* pour trois signaux de parole. L'axe horizontal représente b tandis que l'axe vertical repr'esente la fréquence.

L'algorithme rapide de *QCWT* que nous utilison pour extraire la transformation en ondelettes est décrit en détail dans [44, 48, 46, 49, 55]. 101 niveaux uniformément espacés en fréquence sont obteu par analyse *QCWT* avec 5 niveaux d'octaves, 7 sous-octaves entre chaque octave and une octave supérieure obtenue par interpolation. L'ondelette génératrice d'analyse est définie par:

$$\psi(t) = e^{\frac{-t^2}{16}}(e^{j\,0.69\,\pi\,t} - 4.89098\ 10^{-4}) \tag{17}$$

La projection du signal sur l'espace de fonctions d'échelles n'est pas utilisée. Le signal de parole est échantillonn'e à $\nu_s = 8\ kHz$. Le plan synchrosqueezést sous-échantillonné par 20. Les contributions de chacuns des 20 échantillons sont ajoutées pour la même abscisse b.

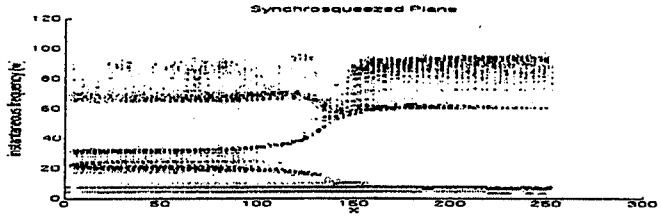


Figure 6: Plan synchrosqueezé pour $/\underbrace{\cdots - a - a}\!-\!\underbrace{i - i - \cdots}/$. Un bruit coloré est présent avec $SNR = 15\ dB$.

Le premier signal est la transition d'une voyelle soutenue $/i/$ à la voyelle soutenue $/a/$, prononcée par une homme dans un environnement bruyant. Le bruits est mesuré sur place à $SNR = 15\ dB$ En partant des fréquences les plus basses, les

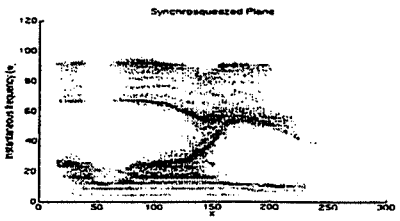


Figure 7: Plan synchrosqueezé pour $/h - \delta - w - a - r - j - u\ ?/$. Un bruit coloré est présent avec $SNR = 15\ dB$.

trois composant primaires rencontrés dans la partie $/a/$ représentent la fréquence fondamentale (pitch) et les deux premiers harmoniques. Ensuite les deux premiers rubans concentrés sont associ'es avec les deux formants. Les deux formants suivants sont également clairement visibles La transition entre $/a/$ et $/i/$ est particulièrement visible et en accord avec ce qui est décrit dans la littérature [21, 10, 67].

Il faut également remarqué l'oscillation du plan synchrosqueezé à la fréquence fondamentale. Les effets de bord n'ont pas été complétement éliminés par *VCE* (V_0 Compatible Extrapolation) strategy décrit dans [48, 46, 44].

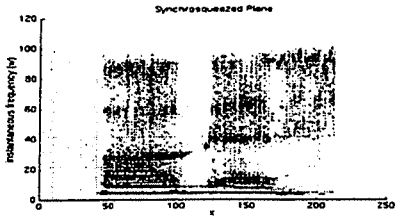


Figure 8: Plan synchrosqueezé pour $/ow - g - \lambda - s - t/$. Un bruit coloré est présent avec $SNR = 15\ dB$.

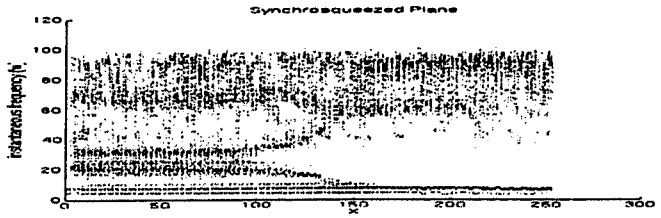


Figure 9: Plan synchrosqueezé pour $/\underbrace{\cdots - a - a}\ - \underbrace{i - i - \cdots}/$. Un bruit coloré est présent avec $SNR = 15\ dB$. Un bruit suplémentaire est ajouté à $SNR = 11dB$

Le second signal est la phrase $/h - \delta - w - a - r - j - u\ ?/$ (How are you?) prononcée avec le même SNR. Le comportment des différent formants est clairement visible, toujours en parfait accord avec la théorie.

Le troisièmem exemple est $/ow - g - \Lambda - s - t/$ (August), qui est partiellement voisée et partiellement non-voisée.

Figure 9 illustre le plan synchrosqueezé pour $/\underbrace{\cdots - a - a}\ - \underbrace{i - i - \cdots}/$ avec un bruit blanc supplémentaire à $SNR = 11dB$.

Une discussion détaillée de param'etres robstes dérivés du plan synchrosqueezé peut être touvé dans [53, 54]. La nature sur-complète de l'analyse et la synthèse pr'esentée dans ce papier contribue énormément à ce comportement robuste [9].

8. ALGORITHME DE FUSION SELECTIVE

Le plan suynchrosqueezé offre lecture immédiate des fréquences instantannées de tous les composants primaires qui ne sont pas masqués. Cependant, il est difficile d'utiliser cette reprsentation pour directement extraire l'amplitude instantannée de ces composants. De plus, il serait avantageux de pouvoir inverser la représentation et de pouvoir reconstruire des portions ou la totalité du plan synchrosqueezé.

8.1 La mesure synchrosqueezée *SFA*

L'algorithme de sélection selective offre cette capabilité [41, 57, 56, 43, 42].

SFA est une autre mesure synchrosqueezée qui, en théorie, extrait automatiquement les lois de modulation de fréquence et d'amplitude de chaque composant primaire. Cependant, en practique, il est généralement implémenté en deux etages successifs:

- Détermination des lois de modulation de féquence à l'aide d'une première mesure synchrosqueezée,
- Extraction de l'amplitude instantannée et de la constante de phase initiale, utilisant l'algorithme *SFA*. Cette étape n'est généralement pas présentée comme étant une mesure synchrosqueezée.

La mesure synchrosquezée non-nécessairement positive associée à *SFA* est

$$\mu_{SFA-aux}(\omega, b, f, g) = \int_{\omega_{WT_f}(a,b,g) = \omega}^{SIS} \int_{-\infty}^{+\infty} WT_f(a, t, g)\ \frac{1}{\sqrt{a}} g\left(\frac{b - t}{a}\right)\ \frac{dt\ da}{C_{g,g}\ a^2} \tag{18}$$

et

$$\mu_{SFA}(\omega, b, f, g) = \sum_k \delta(\omega - \omega_{c_k}) \int_{\omega\ \in\ SqueezedRibbon(\omega_{c_k})} d\mu_{SFA-aux}(\omega, b, f, g) \tag{19}$$

où ω_{c_k} est a fréquence centrale du k^{eme} ruban concentré.

La résistance au bruit est augmentée en implémentant l'int'egrale en a, dans l' équation (18), avec *SIS*.

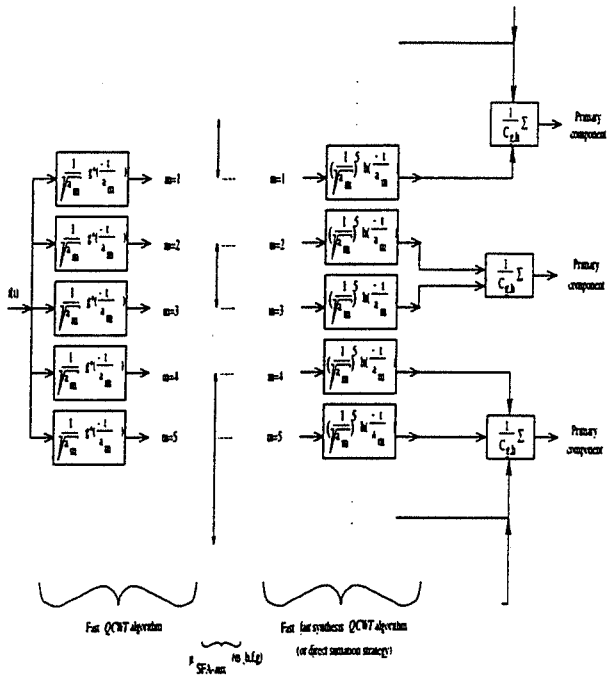


Figure 10: Algorithme de fusion sélective.

SFA produit un plan temps-fréquence où la fréquence centrale de chaque composant concentré suporte le composant primaire reconstruit. Pour ces composants, la mesure synchrosqueezée $\mu_{SFA-aux}(\omega, b, f, g)$ offre lecture immédiate de leurs lois de modalation de fréquence.

8.2 Impl'ementation pratiques

Le principe de base de l'algorithme *SFA* repose sur l'observation de tous les niveaux d'échelles (sous-bandes) de la transformation en ondelettes et sur le regroupement, à un moment donné, des sous-bandes qui présentent une même évolution temporelle d'un propriété [41, 57, 56, 43, 42]. Toutes les sous-bandes qui appartiennent à une même groupe sont reconstruites à l'aide de la formule de reconstruction de l'identité.

Pour des signaux qui satisfont au modèle de modulation, l'algorithme extrait les diff'erent composants primaires. La différence entre le signal initial et la somme des différent composant primaires reconstruits est le signal $\eta(t)$, comme décrit par (1),

$$f(t) - \int_0^{+\infty} \mu_{SFA}(\omega, t, f, g)\, d\omega = \eta(t) \tag{20}$$

Dès lors, si certain composant primaires sont présents mais masqué par d'autres composants dominants, ils peuvent être extraits du reste après extraction des composants dominants.

L'extraction de $\eta(t)$ permets de tester la validité du modèle de modulation pour une application donnée ou pour une famille de signaux. Elle permet également l'évaluation du niveau et du type de bruit présent dans le signal.

Pour des signaux de paroles, ou la validité du modèle de modulation est mieux établie, $\eta(t)$ peut être considéré comme du bruit, du moins pour ds signaux vois'es. En fait, ceci demeure valide our de cortes portions non-voisées entre les portions vois'es. Dès lors, il est possible d'utiliser *SFA* pour débruiter le signal en éliminant tout ce qui bn'a pas un comportement de type "formant".

Il peut également être utilisé comme un classique extracteur de formants avec de nombreuses applications incluant le codage par analyse-synthèse [24, 64, 16].

Rien n'impose d'extraire les composants à regrouper sur la base de la mesure proposée par (18). N'importe quelle autre mesure synchrosqueezée est acceptable.

Dans ce cas plus g'enéral, chaque sous-bande est marqu'ee par le groupe auquel elle appartient, avec la possibilié de n'appartenir à aucun groupe. Pour chaque groupe, à chaque instant, les sous-bandes qui sont marquées comme apartenant à ce groupe sont reconstruites. C'est illustré à la figure 10.

L'algorithme de synthèse, utilis'e pour implementer la résolution de l'identité de chaque groupe doit utiliser équitablement la contribution de chaque sous-bande qui contribue. En $QCWT$, il est souvent possible de simplement utiliser la méthode de sommation directe décrite dans [48, 46]. Dans ce cas, l'équation (18) devient

$$\mu_{SPA-ds}(\omega, b, f, g) = \int_{\omega_{WT_f}(a,b,g) = \omega}^{SIS} WT_f(a, b, g) \frac{dt\, da}{C_{g,\delta}\, a^{\frac{3}{2}}} \tag{21}$$

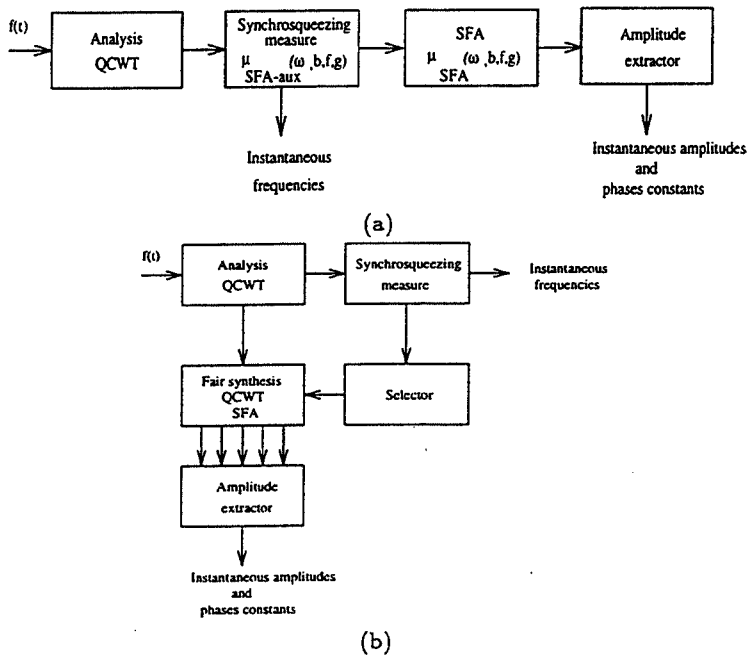


Figure 11: Algorithme d'extraction de la phase et de l'amplitude instantannée. (a) Implémentation théorique (b) Implémentation pratique.

L'algorithme complet d'extraction de l'amplitude et de la phase est schématisé à la figure 11.

Différentes techniques peuvent être utilisées pour extraire l'amplitude instantannée de chaque composant primaire extrait $f_k(t)$. Evidemment, il est possible de directement utiliser la transformation de Hilbert, mais elle n'utilise pas la connaissance de la fréquence instantannée.

Dès lors, une approche plus efficace régresse une fonction d'amplitude pour minimiser la distance $d(.,.)$ [12]:

$$\bar{A}_k(t) = \underset{B(t)}{argmin}\, d(f_k(t), B(t)cos(\Phi_k(t))) \tag{22}$$

Finalement, il est également possible d'adopter le point de vue du modèle de sinusoide [61, 36, 59, 60, 66, 2, 26]: chaque composant primaire peut être décomposé en une série de sinusoides.

8.3 Un extracteur de composants primaires

Les nombreuses oscillations présentes dans le plan synhrosqueezé rendent très difficiles l'extraction rapide et correcte des composants primaires. Un résum'e des difficultés rencontr'ees par des méthodes classiques d'analyse d'images t d'extraction de saliences est pr'esenté dans [48].

Une approche automatique est proposée. C'est une méthode de régression basée sur une approche statistique qui permet une extraction correcte des rubans concentr'es [48, 46, 54].

La méthode proposée dans ce papier est appelée MLE-based living cubic spline extractor (MLE-$LCSE$) [13]. MLE dénote le caractère d'estimation par maximum de vraisemblance [65, 70, 11, 3]. MLE sont bien connus pour sa lente convergence. Notre

[12] Evidement, cela supose que la constante de phase est également connue. Dans le cas contraire, la constante doit est ajustée pour minimiser les distances.

[13] "living" resulte de la possibilité de tuer our créer des splines si n'ecessaire.

approche évite le problème en utilisant les résultats de rapides aggrégations comme premières estimations proches de la solution finale.

Afin d'extraire convenablement les rubans concentrés sur base des amplitudes dominantes dans le plan sychrosqueezé ($\mathcal{V}_{(b,\omega)}$), la mesure synchrosqueezée est interpr'etée comme un fonction de densité de probablilité sur $\mathcal{V}_{(b,\omega)}$. Le problème devient statistique. Dès lors, la densit'e déchantillonnage le long de l'axe temporel est donnée par

$$\rho_s(t) = \int_{\omega \in \mathcal{V}_{(b,\omega)}} d\mu(t,\omega) \tag{23}$$

Initiallement, le modèle suppose N_{pc} composants primaires présents das le plan synchrosqueezé. Le plan $(\omega|t, N_{pc})$, où la notation signifie que les valeurs de ω dépendent de la position t, est modélisé par une combinaison de Gaussiennes,

$$P(\omega|t, N_{pc}) = \sum_{k=1}^{N_{pc}} \frac{\pi_k}{(\sigma_k(t))^2} e^{\frac{-(\omega - \omega_{c_k}(t))^2}{2(\sigma_k(t))^2}} \tag{24}$$

où:

- $P(\omega|t, N_{pc})$ est la probablilité conditionnelle d'observation de ω au temps t, sachant qu'il y a N_{pc} composant principaux.
- $\omega_{c_k}(t)$ est la fréquence centrale du k^{eme} rubban concentré au temps t.
- $(\sigma_k(t))^2$ est la variance du k^{eme} ruban concentré, au temps t. Elles est directeent liée au la largeur de bande du composant primaire asocié.
- π_k est la probilité asocié au k^{eme} ruban concentré.
- $\frac{1}{(\sigma_k(t))^2} e^{\frac{-(\omega - \omega_{c_k}(t))^2}{2(\sigma_k(t))^2}}$ est la probabilité d'observer ω, sachant qu'elle est associée avec le k^{eme} ruban concentré.

La fréquence centrale $\omega_{c_k}(t)$ est décrite par des splines cubiques,

$$\omega_{c_k}(t) = \omega_{c_k}(t, k, \theta_k) \tag{25}$$

où θ_k représente les paramètres définissant la k^{eme} spline cubique. Afin de rester en accord avec la d'efinition de formants, qui ne peuvent pas se croiser, nous interdisons que des composants se croisent. Les paramètres π_k, θ_k and $(\sigma_k(t))^2$ sont estimés pare maximum de vraisemblance [48, 46, 54]

$$\int_{t \in \mathcal{V}_{(b,\omega)}} \rho(t)\, dt\, log_2(P(\omega|t, k, \pi_k, \theta_k, (\sigma_k(t))) \tag{26}$$

Pour la première passe à travers le plan $\mathcal{V}_{(b,\omega)}$, une première estimation est obtenue par aggrégation.

L'équation (26) est résolue en écrivant chaque fréquence centrale comme

$$\omega_{c_k}(t) = \sum_{i=1}^{J_i} \omega_{c_k}(t_i)\, \xi_i(t) \tag{27}$$

où $\xi_i(t)$ sont les fonctions de base d'interpolation par splines cubiques. $\xi_i(t)$ est obtenue en imposant une valeur 1 au temps t_i et zéro à tous les autres t_k. Après avoir estimé les frquence centrales, l'estimation des variances est immédiate.

Après ces estimations la mesure synchrosqueezée de chaque ruban concentré est calculée sur chaque ruban concentré. Cette valeur est comparée aux mesures voisines qui ne sont pas associées avec une ruban concentré ainsi qu'aux valeurs précédentes le long du ruban. Sur la bases de ces valeurs, une décision est prise de conserver ou tuer le ruban.

Figure 12 pr'esente la fréquence centrale obtenue pour $/\underbrace{\cdots - a - a}\underbrace{- i - i - \cdots}/$, bruité.

Une fois extraits, l'algorithme *SFA* peut être utilisé pour sélectionner et reconstruire les différents composants primaires du signal.

L'exposé comparera les paramètres qui peuvent être dérivés des ces composants [53, 54, 50, 40].

9. CONCLUSION

L'exposé compare les performances en reconnaissance de la parole et du locuteur qui peuvent être obtenues en utilisant différents paramètres construits sur le plan synchrosqueezé ou à partir des composants primaires extraits par rapport aux paramètres conventionnels.

En conclusion, le plan synchrosqueezé permet de réarranger le plan temps-fréquence, avec l'avantage supplémentaire d'éliminer la pluspart des perturbations introduites par les outils d'analyse.

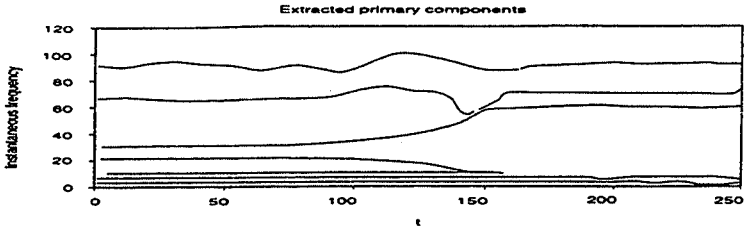


Figure 12: Composants primaires extraits pour $/\underbrace{\cdots - a - a}_{} - \underbrace{i - i - \cdots}_{}/$. Un bruit coloré est pr'esent à $SNR = 15\ dB$. Un bruit blanc supplémentaire est ajouté à $SNR = 10\ dB$.

De plus, l'analyse qui resulte est résistante au bruit. *SFA*, permet de facilement extraire les lois de modulation de f̊quence et d'amplitude des diff'erent composants primaires.

L'approche peut être adapt'ee aux représentations bas'ees sur des mod'eles du nerf auditif. Dès lors, *SFA* permet d'inverser ces repr'esentations.

10. REFERENCES

[1] J. B. Allen. Cochlear modeling. *IEEE ASSP Magazine*, 2(1):pp. 3–29, January 1985.

[2] L. B. Almeida and F. M. Silva. Variable-frequency synthesis: A improved harmonic coding scheme. In *IEEE Proc. ICASSP*, page 27.5.1, San Diego, CA, 1984.

[3] T. Anderson. *An introduction to multivariate statistical analysis.* Wiley, New York, NY, 1958.

[4] K. Assaleh. *Robust features for speaker identification.* PhD thesis, CAIP Center - Rutgers University, The State Univeristy of New Jersey, New Brunswick, NJ, 1993.

[5] B. S. Atal and J. R. Remde. A new model of *LPC* excitation for producing natural-sounding speech at low bit rates. In *IEEE Proc. ICASSP*, pages 614–617, Paris, France, 1982.

[6] F. Auger and P. Flandrin. The why and how of time-frequency reassignment. In *International symposium on time-frequency and time scale analysis*, pages 197–200, Philadelphia, PA, October 1994.

[7] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. on Signal Processing*, May 1995.

[8] J. J. Benedetto and A. Teolist. A wavelet auditory model and data compression. *ACHA*, 1(1):pp. 3–28, 1993.

[9] J.J. Benedetto. Noise reduction in terms of the theory of frames. In *Program on Spline Functions and the Theory of Wavelets*, Montreal, Canada, March 1996. Centre de Recherches Mathématiques, Université de Montréal.

[10] P. Calliope. *La parole et son traitement automatique.* Masson, 1989.

[11] J. V. Candy. *Signal processing: the modern approach.* Mc Graw-Hill, 1988.

[12] J.M. Combes, A. Grossmann, and Ph. Tchamitchian, editors. *Wavelets, time-frequency methods and phase space.* Springer-Verlag, 1989. Proc. of the Int. Conf., Marseille, France, December 14-18, 1987.

[13] E. Copson. *Asymptotic expansions.* Cambridge University Press, Cambridge, MA, 1965.

[14] C. D'Alessandro. Time-frequency speech transformation based on an elementary waveform representation. *Speech Communication*, 9:pp. 419–431, 1990.

[15] I. Daubechies and S. Maes. Wavelets in medicine and biology. In A. Alroubi and M. Unser, editors, *A nonlinear squeezing of the continuous wavelet analysis based on auditory nerve models.* CRC Press, April 1996.

[16] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals.* Macmillan, New York, NY, 1993.

[17] N. Delprat. Extraction of frequency modulation laws in sound synthesis. Research notes in applied mathematics. Masson and Springer-Verlag, 1991.

[18] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, Ph. Tchamitchian, and B. Torrésani. Asymptotic wavelet and gabor analysis: extraction of instantaneous frequencies. *IEEE Trans. on Inf. Theory*, 38(2 Part II):pp. 644–664, 1992.

[19] B. Escudié and B. Torrésani. Wavelet analysis as asymptotic signals (I. Ridge and skeleton of the transform, II. A tentative model for bat sonar receiver). Research notes in applied mathematics. Masson and Springer-Verlag, 1991.

[20] G. Fant. *Acoustic theory of speech production.* Mouton's Co, Hague, 1960.

[21] J. L. Flanagan. *Speech analysis synthesis and perception.* Springer-Verlag, 1972.

[22] D. H. Friedman. Instantaneous-frequency distribution vs. time: an interpretation of the phase structure of speech. In *IEEE Proc. ICASSP*, pages 1121–1124, 1985.

[23] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. ASSP*, 29:pp. 254–272, 1981.

[24] S. Furui. *Digital speech processing, synthesis and recognition.* Marcel Dekker, New York, NY, 1989.

[25] S. Furui and M. Sondhi, editors. *Advances in speech signal processing.* Marcel Dekker, New York, NY, 1991.

[26] O. Ghitza. Auditory nerve representation criteria for speech analysis/synthesis. *IEEE Trans. ASSP*, 6(35):pp. 736–740, June 1987.

[27] O. Ghitza. Advances in speech signal processing. In Furui and Sondhi [25].

[28] O. Ghitza. Auditory models and human performances in tasks related to speech coding and speech recognition. *IEEE Trans. Speech and Audio Proc.*, 2(1):pp. 115–132, January 1994.

[29] C. Gonnet and B. Torresani. Local frequency analysis with two-dimensional wavelet transform. *preprint, CPT-92/P.2870*, January 1993.

[30] S. Greenberg, editor. Representation of speech in the auditory periphery. *Journal of Phonetics*, 16(1), January 1988.

[31] P. Guillemain. Analyse et modélisation de signaux acoustiques. Technical report, CNRS, Laboratoire de Mécanique et d'Acoustique, Marseille, France, 1990. Mémoire de D.E.A. Acoustique et Dynamique des vibrations.

[32] P. Guillemain. Estimation of frequency modulation laws associated to a sum of asymptotic signals with continuous wavelet and gabor transform. *preprint, CPT*, 1992.

[33] P. Guillemain. Ridges associated to isolated frequential and temporal singularities. *preprint, CPT*, 1992.

[34] P. Guillemain, R. Kronland-Martinet, and B. Martens. Application de la transformée en ondelettes à la spectroscopie RMN. Note - laboratoire de mécanique et d'acoustique $n^0$112, Laboratoire de Mécanique et d'Acoustique - CNRS, Marseille, France, June 1989.

[35] P. Guillemain, R. Kronland-Martinet, and B. Martens. Estimation of spectral lines with the help of the wavelet transform, application in NMR spectroscopy. Research notes in applied mathematics. Masson and Springer-Verlag, 1991.

[36] P. Hedelin. A tone-oriented voice-excited vocoder. In *IEEE Proc. ICASSP*, page 205, Atlanta, GA, 1981.

[37] M. Holschneider. General inversion formulas for wavelet transforms. *J. Math. Phys.*, 34(9):pp. 4190–4198, September 1993.

[38] M. Lavrentiev and B. Chabat. *Méthodes de la théorie des fonctions d'une variable complexe.* Mir, Moscow, 1977.

[39] J. S. Liénard. Speech analysis and reconstruction using short-time, elementary, waveforms. In *IEEE Proc. ICASSP*, pages 948–951, Dallas, TX, 1987.

[40] S. Maes. Impact of filtering distorsions on acoustic speech parameters. Viewgraphs presented at NSA, Fort Meade, MD, January 17, 1995.

[41] S. Maes. Quasi-continuous wavelets used to extract parameters in speech processing speech. Viewgraphs presented at the Summer School on Wavelet Techniques and Applications, Bordeaux, France, (September 28, 1993).

[42] S. Maes. Non-linear techniques for parameter extraction from quasi-continuous wavelet transform, with application to speech. In *Proc. EOS-SPIE of the International Symposium on Substance Identification Technologies - Conference on Neural Networks and Artificial Intelligence, paper N^0 2093-03*, Innsbrück, Austria, October 1993. EOS-SPIE.

[43] S. Maes. Application of the wavelet transform for speech processing. In L. Johnson and P. Robinson, editors, *Proc. of the 2nd Annual International Space University Alumni Conference*, NASA Conference Publication 3253, pages 112–120, Huntsville, AL, August 1994. NASA - G. C. M. S. F. C., Marshal Space Flight Center, Alabama. Venue: University of Alabama at Huntville and Marshall Space Flight Center.

[44] S. Maes. A fast quasi-continuous wavelet transform algorithm. In *Proc. Colloque Temps-fréquence, Ondelettes et Multirésolution: Théorie, Modèles et Applications*, Lyon, France, March 1994. Institut National Des Sciences Appliquées de Lyon et CNRS.

[45] S. Maes. *The generalized hybrid wavelet transform yields a new reading of subband coding decompositions.* PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, June 1994. (Annex PhD Dissertation).

[46] S. Maes. The synchrosqueezed plane representation, a new time frequency representation applied to speaker identification. Technical report, CAIP Center - Rutgers University, The State Univeristy of New Jersey, New Brunswick, NJ, June 1994.

[47] S. Maes. Wastrum: wavelet derived generalized cepstrum, using the synchrosqueezed plane. In *Proc. Colloque Temps-fréquence, Ondelettes et Multirésolution: Théorie, Modèles et Applications*, Lyon, France, March 1994. Institut National Des Sciences Appliquées de Lyon et CNRS.

[48] S. Maes. *The wavelet tranform in signal processing, with application to the extraction of the speech modulation model*

features. PhD thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, June 1994.

[49] S. Maes. Fast quasi-continuous wavelet algorithms for analysis and synthesis of 1-D signals. *preprint to appear in SIAM J. Applied Math.*, April 1995.

[50] S. Maes. Investigation of new acoustic features for robust automatic speech recognition. Technical Memorandum, AT&T Bell Laboratories, Speech Research Department, Murray Hill, NJ, July 1995.

[51] S. Maes. LPC Techniques and some new ear-model-based front-ends. Lecture Notes of internal course on Speech Recognition. IBM T. J. Watson Research Center, Human Language Technology, Hawthorne, NY, October 1995.

[52] S. Maes. The synchrosqueezed representation yields a new reading of the wavelet transform. In H. H. Szu, editor, *Proc. SPIE 1995 on OE/Aerospace Sensing and Dual Use Photonics - Wavelet Applications for Dual Use - Session on Acoustic and Signal Processing, Wavelet Applications II*, volume 2491, pages 532–559, Orlando, FL, April 1995. Part I.

[53] S. Maes. The wavelet-derived synchrosqueezed plane representation yields a new time-frequency analysis of 1-D signals, with application to speech. *preprint submitted to ACHA*, April 1995.

[54] S. Maes. The wavelet-derived synchrosqueezed plane representation yields new front-ends for automatic speech recognition. *preprint submitted to IEEE Trans. Speech and Audio Processing*, April 1995.

[55] S. Maes. Signal analysis and synthesis with 1-D quasi-continuous wavelet transform. In *Proc. 12th. International Conference on analysis and optimization of systems*, Paris, June 1996. IRSIA.

[56] S. Maes and I. Daubechies. An application of wavelets to the characterization of formants parameters in speech. Viewgraphs presented at AT&T Bell Laboratories, Holmdel, USA, (May 7, 1993).

[57] S. Maes and I. Daubechies. Wavelets and the modulation model. Technical report, CAIP Center - Rutgers University, The State Univeristy of New Jersey, New Brunswick, NJ, December 1992. in *Speech processing using the modulation model*, presented at Bolling US Air Force Basis, Washington DC, USA, Working Meeting on the Application of Wavelets for Formant Extraction, (December 8, 1992).

[58] B. Martens. Application de l'analyse en ondelettes à la spectroscopie RMN. Technical report, Université Catholique de Louvain, Département de physique, Louvain-la-Neuve, Belgium, 1989. Mémoire de Licence en Physique.

[59] R. J. McAulay and T. F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *IEEE Proc. ICASSP*, page 27.6.1, San Diego, CA, 1984.

[60] R. J. McAulay and T. F. Quatieri. Mid-rate coding based on a sinusoidal representation of speech. In *IEEE Proc. ICASSP*, page 945, Tampa, FL, 1985.

[61] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. ASSP*, 34(4):pp. 744–754, August 1986.

[62] P. M. Morse. *Vibrations and sounds*. Mc Graw-Hill, New York, NY, 1948.

[63] P. M. Morse and K. U. Ingard. *Theoretical acoustics*. Mc Graw-Hill, New York, NY, 1968.

[64] F. J. Owens. *Signal processing of speech*. Macmillan, UK, 1993.

[65] V. Pugachev. *Théorie des probabilités et statistique mathématique*. Mir, Moscow, 1982.

[66] T. F. Quatieri and R. J. McAulay. Speech transformations based on a sinusoidal representation. In *IEEE Proc. ICASSP*, page 489, Tampa, FL, 1985.

[67] L. Rabiner and B-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, Englewoods Cliffs, NJ, 1993.

[68] M. A. Ramalho. *The pitch modulation model with applications in speech processing*. PhD thesis, CAIP Center - Rutgers University, The State Univeristy of New Jersey, New Brunswick, NJ, 1994.

[69] G. Sarraco, G. Sessarego, J. Sageloli, P. Guillemain, and R. Kronland-Martinet. Extraction of modulation laws of elastic shells by the use of the wavelet transform. Research notes in applied mathematics. Masson and Springer-Verlag, 1991.

[70] L. R. Shenton. *Maximum likelihood estimation in small samples*. Macmillan, New York, NY, 1977.

Parole audiovisuelle

Perception visuelle du glide d'arrondissement en français

Marie-Agnès Cathiard

Institut de la Communication Parlée, CNRS UPRESA 5009, INPG-Université Stendhal,
BP 25, F-38040, Grenoble Cedex 9
Tél.: ++04 76 51 76 45 - Fax: ++04 76 82 43 35
e-mail cathiard@icp.inpg.fr

ABSTRACT

Does visual perception of the French rounded glide [ɥ] requires a dynamic representation? For the vowel [y], this is the duration of the static phase of rounding which is necessary to obtain a contrast with the glide. Results with a gating paradigm indicate that the intrinsically dynamic nature of the glide is not exploited until subject perception is oriented towards motion processing. In conclusion we cannot support a visual dynamic representation neither for vowels nor glides.

1. INTRODUCTION

La question de la nature temporelle ou non des représentations de la parole a été posée de longue date en perception auditive. Plusieurs études donnent la faveur à des indices dynamiques pour la perception des voyelles, que ces indices soient dus à la coarticulation des voyelles avec les consonnes ou inhérents à la production des voyelles (p. ex. [Boh95] *vs.* [Nea95]). Plus récemment, la même question a été posée pour la perception visuelle et audiovisuelle. Ainsi, [Ro96] défendent — à l'aide d'expériences testant l'illusion audiovisuelle McGurk par une technique de points lumineux — une option exclusivement cinématique. Cependant des données expérimentales et les résultats en neuropsychologie obtenus par [Ca96a et Ca96b] ne donnent pas la faveur aux seuls mécanismes de traitement du mouvement pour la perception phonétique visuelle et audiovisuelle.

Nos propres données sur la perception visuelle de l'arrondissement en français [Cat94 et Cat96] ne justifient pas non plus un point de vue 'tout-dynamique'. Nous avons en effet proposé, par une approche *shape-from-shading-from-motion*, que la forme de la voyelle serait un possible format représentationnel et le mouvement, seulement une aide pour récupérer la forme lorsque celle-ci est sous-échantillonnée ou non optimalement projetée, comme dans l'arrondissement vu de face — la projection optimale pour le gain de perception de l'anticipation étant de profil. Nous avions cependant émis une réserve à nos conclusions, en signalant le cas particulier des glides qui semblent se différencier des voyelles spécifiquement par leur caractère dynamique ([Kle91]). Les expériences présentées ici apportent des réponses à cette question.

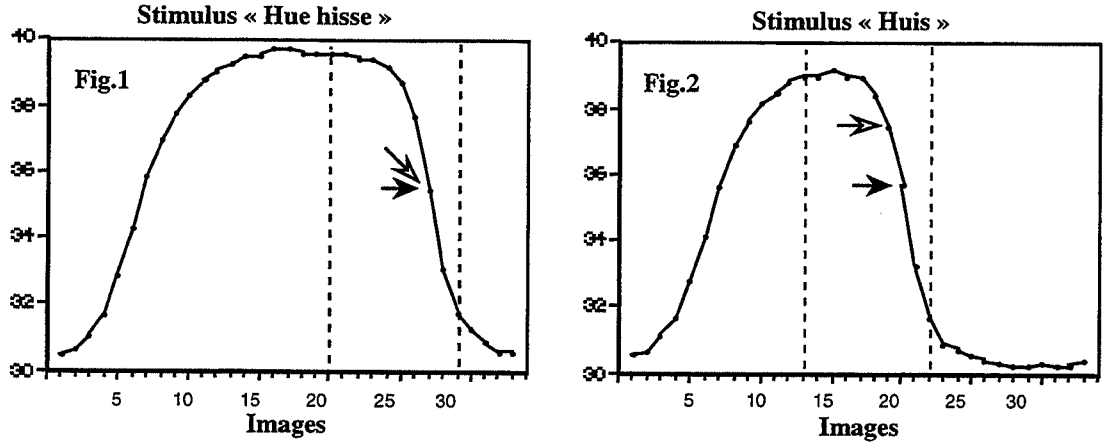
2. STIMULI ET ANALYSE ARTICULATOIRE

Nous avons utilisé des transitions [i # y] et [i # ɥ] insérées dans les phrases porteuses «T'as dis: Hue hisse» [tadi#yis] et «T'as dis: Huis» [tadi#ɥis], où «Hue» et «Huis» sont des noms propres. Ces exemples ont été choisis respectivement pour leurs probabilités de synérèse maximale pour «Huis» et minimale pour «Hue hisse» ([Kle91]).

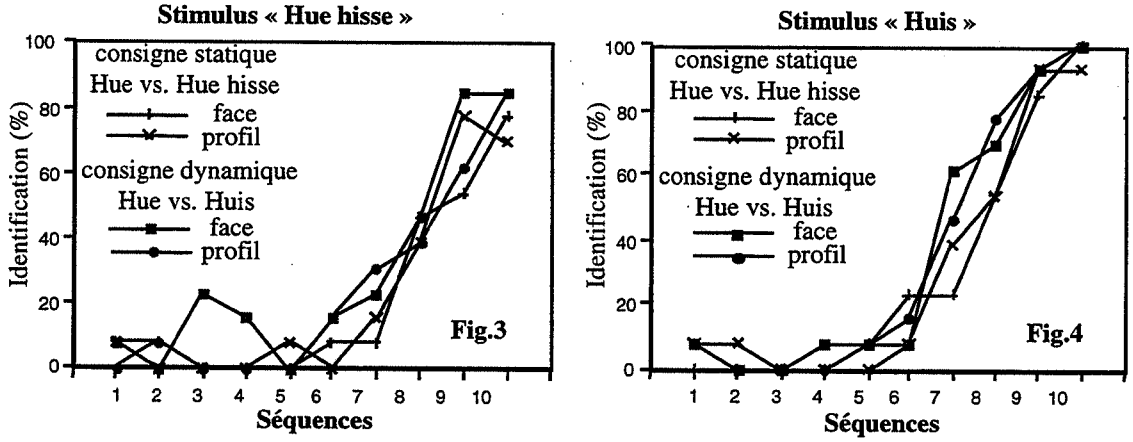
Les deux phrases ont été enregistrées audiovisuellement, à 25 images/seconde, par un locuteur français, simultanément de face et de profil, et répétées 10 fois. Un traitement d'images des stimuli, portant sur plusieurs paramètres aux lèvres [Lal91], a permis de comparer le geste d'arrondissement de la voyelle [y] et celui du glide [ɥ]. Deux séquences explorant l'opposition [ɥ] vs. [y] ont été retenues. Leurs décours articulatoires, comparés sur l'évolution de la protrusion de la lèvre supérieure et de l'aire aux lèvres (qui sont les paramètres les plus pertinents pour ces gestes), se montrent semblables pour les phases d'établissement de l'arrondissement et de rétraction de ce geste. Elles ne diffèrent que sur la phase du plateau de protrusion, qui est nettement de plus longue durée pour le [y] de «T'as dit: Hue hisse» (figures 1 et 2). Notons que ces deux stimuli présentent des valeurs maximales et minimales d'aire aux lèvres et de protrusion très voisines, ainsi que des profils de vitesse et d'accélération très semblables dans les phases de transition.

3. PRÉ-TEST VISUEL

Nous avons monté un prétest visuel pour savoir si les deux séquences «T'as dis: Hue hisse» et «T'as dis: Huis» pouvaient être identifiées en vision seule. Les deux séquences, présentées dans leur totalité, avec un angle de vue de face, ont été répétées 10 fois et présentées en ordre aléatoire. Sur 27 sujets (normalement entendants, francophones, agés de 18 à 20 ans et n'ayant pas de déficit visuel) qui ont participé à ce prétest, 13 sujets seulement ont eu un score individuel d'identification correcte supérieur ou égal à 60% (60% = 1 sujet; 65% = 1; 70% = 3; 75% = 3; 80% = 2; 85% = 1 et 90% = 2 sujets). Ce premier résultat indique qu'il n'est pas si aisé d'identifier, à partir de la seule information visuelle, le glide de sa voyelle correspondante.



Figures 1 et 2: Évolution de la protrusion de la lèvre supérieure (en mm, calculée par rapport à un point de référence sur la règle fixée aux lunettes portées par le locuteur) pour la voyelle [y] de «Hue hisse» et le glide [ɥ] de «Huis». En abcisse est indiqué le nombre d'images : en réalité, il s'agit de trames (une trame paire et une trame impaire composent une image vidéo) qui sont extraites et dont les lignes manquantes sont calculées par interpolation. Les traits verticaux indiquent la zone des images finales des 10 séquences du test 2. La flèche noire situe la frontière d'identification en condition statique et la flèche blanche, celle en condition dynamique.



Figures 3 et 4: Fonctions d'identification du stimulus «Hue hisse» et du stimulus «Huis» en consignes statique et dynamique, de face et de profil. Pour chaque stimulus : en consigne statique, les courbes expriment les pourcentages d'identification «Hue hisse», tandis qu'en consigne dynamique, les courbes expriment les pourcentages d'identification «Huis».

Rappelons que les variétés du français ([Kle91]) ne connaissent pas d'opposition stable sur [y] *vs.* [ɥ], pour des mots comme *muer* vs. *muet*, où l'on rencontre des réalisations [mye], [mɥe]. Cependant, il existe une sensibilité à l'emploi de la synérèse au lieu de la diérèse, comme en témoigne, entre autres, la surprise que peut causer, chez certains locuteurs, le verlan qui traite *puer* comme [ɥɛp] (soit à partir de [pɥe], comme *pied* [pje] —> [jɛp], et non à partir du dissyllabe [pye] —> *[epy]).

4. TEST VISUEL 1 : CONTRIBUTION DE LA DURÉE DU PLATEAU D'ARRONDISSEMENT

Nous avons voulu, dans un premier temps, tester la sensibilité des 13 sujets, qui ont démontré cette compétence phonétique, sur la durée du plateau d'arrondissement en créant des stimuli ayant une durée de plateau intermédiaire entre celle de «T'as dis: Huis» et celle de «T'as dis: Hue hisse». 3 stimuli intermédiaires ont été créés en retirant du stimulus «T'as dis: Hue hisse» successivement une image (au centre du plateau : ce qui correspond sur la figure 1 à la suppression des 2 trames-images n°18 et n°19; = stimulus Hh-1; puisqu'une image vidéo est en réalité composée de deux trames entrelacées), deux images (l'image du centre — n°18 et n°19 — et celle la précédant — n°16 et n°17; = Hh-2) et trois images du plateau d'arrondissement (les 2 précédentes et la suivante — n°20 et n°21; = Hh-3). Ces 3 stimuli ont suffit à créer un continuum entre «T'as dis: Hue hisse» et «T'as dis: Huis» puisque le retrait d'une quatrième image du plateau de «T'as dis: Hue hisse» aurait abouti à un stimulus ayant une durée de plateau égale à la durée de plateau de «T'as dis: Huis».

Nous avons présenté nos 5 stimuli («Huis», «Hue hisse» et les trois stimuli intermédiaires «Hh-1», «Hh-2» et «Hh-3»), avec un angle de vue de face, répétés 10 fois chacun, en ordre aléatoire, en demandant aux sujets de les identifier comme «Huis» ou comme «Hue hisse».

Les pourcentages moyennés d'identification «Hue hisse» sont les suivants : 18.5% pour «Huis», 28.5% pour «Hh-1», 44.6% pour «Hh-2», 69.2% pour «Hh-3» et 80% pour «Hue hisse». Ainsi, plus la durée du plateau augmente, plus les identifications «Hue hisse» augmentent.

Nous pouvons conclure de cette expérience que la durée du plateau d'arrondissement semble être un bon indice visuel pour distinguer le glide arrondi [ɥ] de sa voyelle correspondante [y]. Cependant, la présentation des phrases dans leur totalité ne nous permet pas de connaître *à partir de quel moment* les sujets ont suffisamment d'information pour décider qu'ils sont en présence du glide ou de la voyelle.

5. TEST VISUEL 2 : CONTRIBUTION DU TIMING DE L'APPARITION DE LA RÉTRACTION

Si la durée du plateau d'arrondissement est un bon indicateur pour la réalisation de la tâche demandée, il est probable que les sujets doivent attendre la fin de ce plateau, autrement dit le début de la phase de rétraction de l'arrondissement, pour évaluer cette durée. Nous avons donc décidé d'explorer le décours temporel de cette phase de rétraction au moyen d'une technique qui nous donnera pas à pas l'évolution de l'identification, soit le *fenêtrage* (ou *gating*).

Pour chaque séquence «T'as dis: Huis» et «T'as dis: Hue hisse», nous avons préparé 10 fenêtres de 1200 ms qui comprenaient toujours le début de la phrase jusqu'au milieu du plateau d'arrondissement. La fenêtre glisse ensuite par pas de 20 ms le long du retrait du geste (cf. figures 1 et 2; le pas de 20 ms est possible en extrayant l'une des deux trames de l'image, dont on reconstitue par interpolation les lignes manquantes). Nous avons en outre décidé de tester les identifications sous deux angles de vue : de face et de profil (l'arrondissement des voyelles étant au mieux récupéré sous ce dernier angle, cf. [Cat94 et Cat96]). Nous avons ainsi au final 4 tests : «Hue hisse» face, «Hue hisse» profil, «Huis» face et «Huis» profil, au sein desquels les 10 séquences tronquées sont présentées en ordre aléatoire.

En plus de cette procédure de fenêtrage, nous avons manipulé l'*attente* de nos sujets. Nous sommes partis du fait que tant que le sujet ne voit que la phase de plateau, il s'attend plus probablement à ce que la voyelle soit [y]. Ce n'est qu'avec l'amorce du geste de rétraction, qui lui signale la fin du plateau, qu'il pourra décider s'il est en présence d'un [y] ou d'un [ɥ]. La manipulation de la consigne vise à modifier chez nos sujets l'attente du moment où l'événement de rétraction est susceptible de se produire. Chacun des 4 tests a été ainsi présenté deux fois aux sujets. Une fois en donnant une consigne dite «statique», où les sujets ont à identifier le stimulus tronqué comme «T'as dit: Hue» ou comme «T'as dit: Hue hisse». Et une autre fois avec une consigne dite «dynamique», où les sujets ont à identifier le même stimulus tronqué comme «T'as dit: Hue» ou comme «T'as dit: Huis» (les stimuli, l'ordre des consignes et l'angle de vue ont été contrebalancés). Rechercher un tel effet de consigne revient à se demander si l'on tire davantage parti du mouvement quand on s'attend à une configuration aux lèvres plutôt dynamique (cas de la consigne dite «dynamique» où les sujets sont préparés à l'identification possible d'un glide) que lorsque l'on s'attend à une configuration plutôt stable (cas de la consigne dite «statique» où les sujets sont préparés à percevoir une voyelle).

Nous présentons les courbes d'identification obtenues, pour les 13 sujets, pour «Hue hisse» (figure 3) et «Huis» (figure 4), de face et de profil, en consignes statique et dynamique. Toutes les courbes d'identification présentent une forme classique en S. Notre intérêt sera spécifiquement centré sur les frontières, classiquement à 50%.

En ce qui concerne le stimulus «Hue hisse» (figure 3), on peut observer une stabilité des frontières à 50% qui se situent, pour les quatre courbes, à l'image finale de la séquence 8, ce qui correspond sur nos tracés articulatoires (figure 1), à un moment où la rétraction est déjà bien amorcée (presque à la moitié de son décours).

En ce qui concerne le stimulus «Huis» (figure 4), on peut observer qu'en consigne statique, de face comme de profil, les frontières à 50% se situent elles aussi à l'image finale de la séquence 8. En consigne dynamique, pour les deux angles de vue, les frontières se situent au niveau de l'image finale de la séquence 7 et sont donc plus précoces de 20 ms. Dans cette condition de consigne, la rétraction est donc prise en compte plus tôt par nos sujets, à un moment où le geste de rétraction en est à son début (figure 2).

En résumé, on n'observe pas de différence de frontière selon l'angle de vue. Les frontières observées pour le stimulus [y] de «Hue hisse», en consignes statique et dynamique, correspondent aux frontières du stimulus glide [ɥ] de «Huis» en consigne statique. Par contre, la frontière perceptive visuelle est plus précoce de 20 ms pour le stimulus glide [ɥ] en consigne dynamique par rapport à sa frontière en consigne statique.

Les sujets ne semblent ainsi pas traiter différemment sur le plan visuel le décours de la voyelle [y] et celui du glide [ɥ], *sauf dans le cas où leur attente est particulièrement centrée sur la précocité possible de l'apparition de la rétraction.* Au vu de ce patron de résultats, on peut dire que pour tirer avantage de la nature plus dynamique d'un stimulus de parole visuelle, il faut : (i) que le stimulus s'y prête (dans notre cas, quand le sujet a un glide à traiter); (ii) mais aussi que l'on soit dans une *attente* (comme c'est le cas avec la consigne dite dynamique) propice au traitement d'événements, pour leurs propriétés cinématiques ou métriques (rythmiques).

6. CONCLUSION

Avec ces premiers résultats sur la perception visuelle du glide d'arrondissement en français, nous avons pu montrer que la moitié seulement de nos sujets (13 sur 27) étaient capables d'identifier *visuellement* le glide par rapport à la voyelle. Pour ces 13 sujets, nous avons mis en évidence que la durée du plateau d'arrondissement — que nous avons fait varier en la raccourcissant — permet bien de passer de la perception de la voyelle [y] à la perception du glide [ɥ]. Nous pouvons comparer ce résultat à ceux de [Gre95] sur la perception visuelle de consonnes plosives simples et géminées en allemand, qui montrent que la variation de la durée de l'occlusion, ou phase statique, est pertinente, très nettement pour les plosives visibles comme [p] (moins pour [t]).

Concernant la question des formats représentationnels pour les voyelles et les glides, nous avons pu montrer que les sujets testés ne semblent pas traiter différemment la cinématique du glide [ɥ] et celle de sa voyelle correspondante [y]. Ce n'est, en effet, que lorsque leur perception est guidée, par une manipulation de l'*attente*, vers la composante plus dynamique du glide que l'information de mouvement offre un avantage. Il resterait à évaluer comment ces résultats obtenus par une perception orientée vers le traitement du mouvement sont compatibles avec la théorie de l'«expectancy» de Jones ([Jon89]), qui semble rester féconde dans la perception des rythmes et des flux musicaux.

Plus spécifiquement, nous ne défendons toujours pas un point de vue 'tout-dynamique' pour la perception *visuelle* des voyelles et des glides, même s'il reste néanmoins possible que, du point de vue acoustique et bimodal, les glides requièrent une représentation intrinsèquement dynamique pour se différencier des voyelles correspondantes.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Boh95] Bohn O.S. & Strange W. (1995), "Discrimination of coarticulated German vowels in the silent-center paradigm : 'Target' spectral information non needed", XIIIth Int. Congr. of Phonetic Sciences, Vol. 2, pp. 270-273, Stockholm.

[Ca96a] Campbell R. (1996), "Seeing brains reading speech : A review and speculations", In D. Stork & M. Hennecke (Eds.), Speechreading by Humans and Machines, NATO ASI Series F : Computer and Systems Sciences, Vol. 150, pp. 115-133, Springer-Verlag.

[Ca96b] Campbell R. (1996), "Seeing speech in space and time: Psychological and neurological findings", In Proceedings of the Fourth International Conference on Spoken language Processing, Oct. 3-6 1996, Philadelphia, PA, USA, pp. 1493-1496.

[Cat94] Cathiard M.-A. (1994), "La perception visuelle de l'anticipation des gestes vocaliques : Cohérence des événements audibles et visibles dans le flux de la parole", Thèse de Psychologie Cognitive, Grenoble 2.

[Cat96] Cathiard M.-A., Lallouache M.-T. & Abry C. (1996), "Does movement on the lips mean movement in the mind?", In D. Stork & M. Hennecke (Eds.), Speechreading by Humans and Machines, NATO ASI Series F, Vol. 150, pp. 211-219, Springer-Verlag.

[Gre95] Greisbach R. (1995), "Parameterization and integration of facial speech gestures", Rapport du Projet Speech Maps. Mapping of Action and Perception in Speech. Esprit/BR N°6975, Year 3, Vol. IV et V, Deliverable 33, RP2, pp. 23-38.

[Jon89] Jones M.R. & Boltz M. (1989), "Dynamic attending and responses to time", Psychological Review, 3, pp. 459-491.

[Kle91] Klein M. (1991), "Vers une approche substantielle et dynamique de la constituance syllabique. Le cas des semi-voyelles et des voyelles hautes dans les usages parisiens", Thèse de Doctorat (nouveau régime), Saint-Denis, Université Paris 8.

[Lal91] Lallouache M.-T. (1991), "Un poste 'visage-parole' couleur. Acquisition et traitement automatique des contours des lèvres", Thèse de l'ENSERG, Spécialité: Signal, Image, Parole, Grenoble.

[Nea95] Nearey T.M. (1995), "Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English", XIIIth Int. Congr. of Phonetic Sciences, Vol. 2, pp. 678-681, Stockholm.

[Ros96] Rosenblum L.D. & Saldaña H.M. (1996), "An audiovisual test of kinematic primitives for visual speech perception", Journal of Experimental Psychology: Human Perception and Performance, 22(2), pp. 318-331.

Remerciements. À Tahar Lallouache avec l'aide duquel nous avons réalisé les expériences rapportées ici. À Jean-Luc Schwartz, notre locuteur. Au même et à Christian Abry pour leurs commentaires fructueux. Les recherches exposées ont été réalisées dans le cadre d'un projet Région Rhônes-Alpes (G. Tiberghien, dir., Institut des Sciences Cognitives, Lyon).

Interactions audiovisuelles dans la perception de la parole en français

C.Colin (1), M. Radeau (2) & P. Deltenre (1)
(1) Université libre de Bruxelles, Belgique; (2) F.N.R.S., Bruxelles, Belgique
Laboratoire de Psychologie Expérimentale
Université libre de Bruxelles CP191
50, av. F. Roosevelt 1050 Bxl
Tél.: ++32.2.650.45.49 - Fax: ++32.2.650.22.09
e-mail: ccolin@ulb.ac.be

Abstract

The present experiment examined the conditions that elicit illusions of the McGurk type in French. We manipulated items length (monosyllables vs bisyllables), the nature of the vowel (/i/ vs /a/), the voiced-voiceless characteristic of some stop consonants and, in the case of the bisyllables, the articulation rate (slow, normal or fast). While many combinations occurred, perceptual fusions were very scarce. The effects were not affected by syllabic length but, for bisyllables, they increased as articulation rate decreased. The vowel /i/ gave rise to more illusions than /a/. Voiceless consonants produced more combinations than voiced consonants, but for fusions there was a small tendency toward the reverse pattern. The results are discussed in terms of the possibility that different mechanisms underly the two kinds of illusions.

1. Introduction

Les interactions audiovisuelles qui se manifestent dans la parole sont illustrées par les illusions qui se produisent en cas de conflit entre la parole auditive et la lecture labiale. Deux types d'illusion ont été observés: des fusions et des combinaisons [McG76]. Ainsi par exemple, lorsque le son /b/ est présenté en même temps que l'image vidéo de la bouche d'un locuteur qui prononce /g/ une fusion peut se produire se manifestant par la perception de /d/. Par contre, la présentation inverse peut donner lieu à une combinaison qui se manifeste par la perception de /bg/. La perception linguistique ne paraît donc pas exclusivement auditive ou visuelle mais semble résulter d'une intégration des deux types d'information.

L'illusion McGurk a été considérée par les protagonistes de la théorie motrice de la perception du langage [Lib85] comme supportant l'idée d'un module perceptif spécialisé, indépendant du module auditif général. Selon d'autres théories, cependant, les interactions audiovisuelles qui interviennent dans la parole ou dans l'analyse de la scène spatiale reposeraient sur les mécanismes perceptifs généraux [Mas87]. En accord avec l'hypothèse modulaire, l'effet McGurk a été décrit comme un phénomène obligatoire, irrépressible, peu influencé par nos connaissances. Les arguments systématiques en faveur de la notion d'impénétrabilité cognitive dans la parole audiovisuelle ne sont cependant pas tout à fait cohérents. Le simple rôle d'une manipulation cognitive consistant en une discordance de genre visage-voix a donné lieu à des résultats contradictoires [Gre91; Wal95], et il en est de même de manipulations d'attention par le biais d'instructions [Eas82; Sum84]. D'autre part, l'illusion est d'autant plus importante que le locuteur articule clairement et lentement [Mun96]. En outre, l'universalité du phénomène ne semble pas clairement établie. La plupart des travaux qui ont donné lieu à des effets importants ont été menés en anglais [Mac78; McG76; Mun96] et quelques-uns en néérlandais [Vro92], allemand [Die95] ou finlandais [Sam97]. Par contre, en japonais, l'effet s'est avéré très difficile à obtenir, n'apparaissant que dans des conditions d'écoute dégradées [Sek91; Sek93]. L'expérience présente a eu pour but principal d'étudier l'effet McGurk en français et plus précisément de mettre en évidence les conditions les plus favorables à son obtention. Nous avons manipulé différentes variables susceptibles d'affecter l'illusion: la longueur des items (une syllabe de type VC vs. deux syllabes de type VCV), le débit des bisyllabes (lent vs. normal vs. rapide), la nature de la voyelle (/a/ vs. /i/) et le caractère voisé ou sourd de consonnes occlusives (/b/ vs. /p/ et /g/ vs. /k/). En ce qui concerne le nombre de syllabes, certains auteurs [Mac78; McG76] utilisent des monosyllabes, d'autres [Mun96] des bisyllabes, mais il ne semble pas y avoir eu de comparaison systématique. L'utilisation de bisyllabes permet de faire varier la vitesse d'articulation laquelle semble aussi avoir été rarement considérée si ce n'est par Munhall & Jones [Mun96] qui comme on l'a vu ont obtenu un effet du débit. L'environnement vocalique ne semble pas avoir été souvent manipulé; Green et al. [Gre88], cependant ont montré que la voyelle /i/ produisait davantage d'effet que la voyelle /a/. Certaines données suggèrent une influence possible du voisement. Ainsi, bien que la grande majorité des chercheurs utilisent uniquement des consonnes voisées, McGurk & Mac Donald [Mac78; McG76] ont utilisé des consonnes sourdes (/p/ et /k/) et voisées (/b/ et /g/) et ont obtenu plus de fusions pour les voisées. En raison de la modulation possible des effets par des paramètres acoustiques (voix grave vs. aiguë) ou physiques (taille de l'ouverture de la bouche), l'influence éventuelle du genre du locuteur ne pouvait être exclue ce qui nous a amené à utiliser deux locuteurs de genre différent.

2. METHODOLOGIE

2.1 Participants

36 étudiants (19 femmes et 17 hommes) ont participé à l'expérience dans le cadre de leurs travaux pratiques en Psychologie. Ils étaient tous âgés de 18 à 25 ans, francophones et sans problème d'audition connu. Leur vision était parfaite ou corrigée.

2.2 Matériel

Les syllabes ont été prononcées par deux locuteurs, un homme (24 ans) et une femme (35 ans). Il s'agissait de monosyllabes de type CV (ba, ga, pa, ka, bi, gi, pi et ki) et de bisyllabes de type VCV (aba, aga, apa, aka, ibi, igi, ipi et iki). Ces dernières ont été prononcées selon trois débits: normal, lent et rapide. Afin d'attirer davantage l'attention des participants sur les lèvres des locuteurs, seule la partie inférieure du visage de ces derniers a été filmée (du sommet du nez au bas du menton).

Tous les items ont été construits sur banc de montage Panasonic. Les items expérimentaux étaient tous incongruents. Nous les avons montés en plaçant, sur le canal audio d'une syllabe donnée, le son correspondant à une autre syllabe. Les deux items dont nous intervertissions les bandes sonores avaient toujours la même longueur, à 2 ou 3 images près (une image=40 ms). Le son était synchronisé sur la dernière image précédant l'ouverture des lèvres. Huit types de stimuli ont été construits pour chaque longueur d'items et dans le cas des bisyllabes pour chaque débit (par exemple, aba a été couplé avec aga, apa avec aka, ibi avec igi et ipi avec iki). Chaque syllabe était tour à tour présentée en modalité auditive ou visuelle. Les items contrôles étaient constitués des mêmes syllabes auditives, mais celles-ci ont été simplement copiées sur une période de visage immobile. Tous les items avaient une durée moyenne de 2 à 3 secondes. Quatre cassettes ont été montées; deux par locuteur, l'une comportant les items expérimentaux et l'autre les items contrôles. Chaque cassette était constituée de 4 blocs, correspondant chacun à l'un des trois débits des bisyllabes et aux monosyllabes. Chaque bloc comprenait 24 essais (8 types d'essais différents répétés 3 fois) présentés dans un ordre aléatoire et 2 ou 3 essais "attrapes" que les sujets devaient détecter (il s'agissait de grimaces telles un penchement ou un hochement de tête, ou encore une protrusion de la langue). Ces "attrapes" étaient destinées à s'assurer que les participants regardaient effectivement l'écran.

2.3 Procédure

Les sujets étaient assis devant une table, à 75 cm d'un écran couleur Panasonic de 39 cm de diagonale. Un haut-parleur Trust était placé au-dessus de l'écran. L'intensité sonore était de 70 dB en moyenne. Les participants recevaient pour consigne d'entourer sur un protocole la réponse correspondant à ce qu'ils avaient entendu. Ils avaient le choix entre /b/, /g/, /p/, /k/, /d/, /t/, /bg/, /pk/. ou "autre". L'intervalle inter-stimuli était de trois secondes. La session commençait par 20 essais d'entraînement. L'ensemble de la session expérimentale durait approximativement une heure.

3. RÉSULTATS

Pour chaque item, les sujets étaient susceptibles de donner une réponse correspondant à l'information auditive, visuelle, ou audiovisuelle (une combinaison ou une fusion). Nous avons considéré comme réponses illusoires les réponses audiovisuelles et visuelles. Bien que les essais contrôles aient, dans plus de 90% des cas, été correctement identifiés, les erreurs à ces stimuli ont été soustraites des totaux lorsqu'elles consistaient en la réponse attendue en cas de combinaison ou de fusion (par exemple, la réponse /bg/ ou /d/ au son /b/).

Table 1: pourcentages de fusions et de combinaisons chez les deux locuteurs, pour les monosyllabes et pour les bisyllabes dans les trois conditions de débit (L=lent; N=normal; R=rapide) et en moyenne (M).

	Combinaisons					Fusions				
	Bisyllabiques				Mo	Bisyllabiques				Mo-
	L	N	R		no	L	N	R	M	no
Locuteur masculin										
bg/a	23	31	17	24	29	3	5	3	4	3
bg/i	59	31	38	43	10	2	10	3	5	0
pk/a	57	20	78	52	39	2	2	4	3	2
pk/i	53	47	36	45	44	3	5	3	4	1
Tot	**48**	**32**	**42**	**41**	**31**	**3**	**6**	**3**	**4**	**2**
Locuteur féminin										
bg/a	46	44	31	40	33	4	3	4	4	2
bg/i	37	54	34	42	47	19	1	2	7	0
pk/a	40	59	16	38	47	7	3	3	4	0
pk/i	66	49	26	47	64	1	2	1	1	0
Tot	**47**	**52**	**27**	**42**	**48**	**8**	**2**	**3**	**4**	**1**

Nous avons effectué huit analyses de variance séparées pour les deux types d'illusion, les deux longueurs syllabiques et les deux locuteurs. Les résultats des deux locuteurs ont été traités séparément car pour un même débit, le locuteur masculin était systématiquement plus rapide que le féminin. Les facteurs inclus dans les analyses étaient: le type de consonnes, le type de voyelles et, dans le cas des bisyllabes, le débit.

3.1 Items bisyllabiques

Combinaisons

En ce qui concerne le locuteur masculin, tous les effets principaux sont significatifs. La voyelle /i/ donne lieu à 7% d'illusions de plus que la voyelle /a/ ($F(1,35)=5,25$, $p<.05$). Les consonnes non voisées donnent 15% de combinaisons de plus que les consonnes voisées ($F(1,35)=18,58$, $p<.01$). L'effet du débit ($F(2,70)=3,14$, $p<.05$) est dû au fait que le débit lent donne 16% de combinaisons de plus que le débit normal ($F(1,35)=6,59$, $p<.01$). La condition de débit rapide ne se différencie pas significativement des deux autres. Nous observons également une interaction entre le type de consonne et le type de voyelle ($F(1,35)=7,00$, $p<.01$), le débit et le type de consonne ($F(2,70)=6,17$, $p<.01$) et une triple interaction entre le type de voyelle, le type de consonne et

le débit ($F(2,70)=18,56$, $p<.01$). Au sein de ces trois interactions, tous les contrastes significatifs concernant la voyelle vont dans le même sens que l'effet principal (/i/ donne plus d'illusions que /a/) à une exception près. L'effet de la consonne est toujours congruent avec l'effet principal, les consonnes non-voisées donnant plus de combinaisons que les voisées. Quant à l'effet du débit, il va dans la plupart des cas dans le sens d'une supériorité du lent sur le normal.

Chez le locuteur féminin, le débit est le seul effet principal significatif ($F(2,70)=11,40$, $p<.01$). Les débits lent et normal donnent respectivement lieu à 20% et à 25% de combinaisons de plus que le débit rapide (lent: $F(1,35)=11,75$, $p<.01$; normal: $F(1,35)=15,97$, $p<.01$). L'interaction entre type de consonne et débit est significative ($F(2,70)=5,38$, $p<.01$) ainsi que la triple interaction ($F(2,70)=8,51$, $p<.01$). L'effet de la consonne va dans le même sens que chez le locuteur masculin: les consonnes non-voisées donnent généralement lieu à plus de combinaisons que les voisées.

Fusions

Aucun effet significatif n'est observé chez le locuteur masculin. Chez le locuteur féminin, cependant, l'effet du type de consonne est significatif: les voisées donnent lieu à 3% de plus de fusions que les non-voisées ($F(1,35)=7,00$, $p<.01$). L'effet du débit est également significatif ($F(2,70)=7,83$, $p<.01$), le débit lent donnant 6% de fusions de plus que le débit normal ($F(1,35)=9,03$, $p<.01$) et 5% de plus que le débit rapide ($F(1,35)=7,52$, $p<.01$). Toutes les interactions doubles sont significatives (Voyelle x Consonne: $F(1,35)=5,31$, $p<.05$; Voyelle x Débit: $F(2,70)=4,86$, $p<.05$; Débit x Consonne: $F(2,70)=5,19$, $p<.01$) de même que la triple interaction ($F(2,70)=7,83$, $p<.01$). Dans tous les cas de contrastes significatifs, les fusions restent plus nombreuses pour les consonnes voisées que pour les non-voisées et dans la condition de débit lent que pour les autres débits. Bien que l'effet principal de la voyelle ne soit pas significatif, la tendance observée précédemment (illusions plus nombreuses pour /i/ que pour /a/) est observée dans le cas du débit lent et de consonnes voisées.

3.2 Items monosyllabiques

Combinaisons

Seul l'effet de la consonne est significatif chez le locuteur masculin, les consonnes non-voisées donnant à nouveau 22% de combinaisons de plus que les non-voisées ($F(1,35)=13,65$, $p<.01$). Un effet analogue de la consonne est aussi présent dans le cas du locuteur féminin ($F(1,35)=11,17$, $p<.01$), chez qui l'effet de la voyelle est également significatif, les /i/ suscitant plus de combinaisons que les /a/ ($F(1,35)=6,66$, $p<.01$).

Fusions

Aucun effet principal, ni aucune interaction, n'a été mis en évidence pour aucun des deux locuteurs.

3. 3 Comparaison des items monosyl- labiques et bisyllabiques

Les résultats des monosyllabes ont été comparés aux résultats moyens obtenus dans les trois conditions de débit des bisyllabes dans quatre analyses de variance (deux locuteurs et deux types d'illusions) à trois facteurs (consonne, voyelle, longueur syllabique). Dans aucune des analyses, un effet de longueur syllabique n'a été observé. Dans l'ensemble, les effets obtenus pour la voyelle et la consonne confirment les tendances observées précédemment.

4. DISCUSSION

L'asymétrie observée ici entre les deux formes de l'illusion McGurk constitue sans doute le point essentiel des résultats. Alors que la présentation d'une occlusive labiale visuelle en même temps qu'une occlusive non labiale (vélaire) auditive donne lieu à des combinaisons perceptives dans 41% des cas, la présentation inverse ne donne lieu à une illusion de fusion que dans 3% des essais. Le pourcentage de combinaisons n'est pas très différent de celui obtenu dans les premières études de McGurk & McDonald [McG76; Mac78]. Par contre, le nombre extrêmement restreint de fusions dans notre étude est très difficile à concilier avec les pourcentages élevés obtenus par ces auteurs (de 35% à 90% selon les conditions). La distinction combinaisons-fusions n'est faite que dans certains travaux. Ainsi, elle n'est pas faite en Japonais où des effets très faibles ont été rapportés [Sek91]. Comme on l'a vu précédemment, la dégradation des conditions d'écoute a permis d'obtenir l'effet McGurk en japonais. Le degré d'intelligibilité du son, de même que le niveau d'intensité sonore, ne sont pas non plus spécifiés dans toutes les études. C'est le cas dans celles de McGurk & McDonald citées plus haut. Notons que Diesch [Die95] dans des conditions d'écoute assez semblables à celles de notre situation tant par l'absence de bruit surimposé au son que par le niveau d'intensité sonore proche de 70 dB a obtenu en allemand des pourcentages très faibles d'illusion: 4% à 9% de combinaisons et 8% à 17% de fusions avec présentation des stimuli visuels dans l'un ou l'autre hémichamp.

Nos résultats concernant l'environnement vocalique confirment ceux obtenus par Green et al. [Gre88]. Quels que soient le locuteur et la longueur syllabique, la voyelle /i/ a toujours tendance a produire plus d'illusions que la voyelle /a/. Ce résultat peut être mis en rapport avec l'intensité plus faible de la voyelle /i/ que de la voyelle /a/ [Ros71]. Etant donné la relation bien établie entre l'illusion et l'intensité des stimulations auditives, la différence d'intensité entre /a/ et /i/ mesurée dans notre situation (9 dB chez le locuteur féminin et 4 dB chez le locuteur masculin) est sans doute à la base de cet effet. Les consonnes non-voisées donnent lieu à plus de combinaisons que les consonnes voisées. Par contre, lorsqu'il s'agit des fusions, ce sont les consonnes voisées qui ont tendance à susciter le plus d'illusions. Cette

asymétrie devrait être réenvisagée en conditions d'écoute dégradée afin de préciser la relation entre la sensibilité aux confusions acoustiques dans le bruit des occlusives utilisées ici [Mil55] et leur susceptibilité à l'influence visuelle. Aucun effet de longueur syllabique n'a été mis en évidence mais dans le cas des bisyllabes un effet important du débit est apparu: le débit lent provoque plus d'illusions que le débit normal, qui lui-même en suscite davantage que le débit rapide. Cette observation est congruente avec les résultats de Munhall & Jones [Mun96]. Le fait que le pourcentage d'illusion augmente à mesure que le débit ralentit suggère une contribution de facteurs stratégiques ce qui ne plaide pas en faveur d'une description de l'effet McGurk en tant que phénomène entièrement modulaire [Rad94a; Rad94b].

La grande discordance entre le pourcentage de fusions et de combinaisons ainsi que l'effet différentiel du type de consonnes sur ces deux illusions suggèrent que les mécanismes à la base de ces deux effets pourraient être différents. Diesch [Die95] avait déjà émis cette hypothèse après avoir mis en évidence un avantage des fusions dans l'hémisphère droit et des combinaisons dans l'hémisphère gauche du cerveau. Cependant, cet auteur ne fournit pas d'argument clair permettant de comprendre pourquoi l'un ou l'autre effet serait plus latéralisé à droite ou à gauche. En outre, il observe, dans la même étude portant sur l'allemand, un pourcentage global de fusions plus élevé que celui des combinaisons. Les illusions McGurk pourraient donc être modulées en fonction de la langue utilisée. En effet, en anglais, elles sont décrites comme très robustes [McG76; Mac78; Mun96]. En japonais, elles paraissent difficiles à obtenir [Sek91; Sek93]. En allemand, les fusions seraient obtenues en plus grand nombre que les combinaisons et en français, l'effet inverse semble se produire. Les mécanismes d'intégration audiovisuels à la base des combinaisons et des fusions pourraient donc non seulement différer entre eux, mais aussi différer en fonction de la langue. Des études ultérieures utilisant des conditions d'écoute dégradée sont nécessaires afin de vérifier les tendances observées ici, en particulier afin de déterminer si le faible pourcentage de fusions est lié à nos conditions expérimentales ou est spécifique du français et si les différences liées au caractère voisé ou non des occlusives utilisées persistent entre fusions et combinaisons.

BIBLIOGRAPHIE

[Die95] Diesch E. (1995) "Left and right hemifield advantages of fusions and combinations in audiovisual speech perception", The Quarterly Journal of Experimental Psychology, 48A, 320-333.

[Eas82] Easton R.D. & Basala M. (1982) "Perceptual dominance during lipreading", Perception and Psychophysics, 32, 562-570.

[Gre88] Green K.P., Kuhl P.K. & Meltzoff A.N. (1988) "Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment", Journal of the Acoustical Society of America, 84, S155.

[Gre91] Green K.P., Kuhl P.K., Meltzoff, A.N. & Stevens, E.B. (1991) "Integrating speech information across talkers, gender and sensory modality: Female faces and male voices in the McGurk effect", Perception and Psychophysics, 50, 524-536.

[Lib85] Liberman A.M. & Mattingly I.G. (1985) "The motor-theory of speech revised", Cognition, 21, 1-36.

[Mac78] MacDonald J. & McGurk H. (1978) "Visual influences on speech perception processes", Perception and Psychophysics, 24, 253-257.

[Mas87] Massaro D.W. (1987), Speech Perception by Ear and Eye: A Paradigm for Psychosocial Inquiry, Erlbaum.

[McG76] McGurk H. & MacDonald J. (1976) "Hearing lips and seeing voices", Nature, 264, 746-748.

[Mil55] Miller G.A. & Nicely P.E. (1955) "An analysis of perceptual confusions among some English consonants", Journal of the Acoustical Society of America, 27, 338-352.

[Mun96] Munhall K.G., Gribble P., Sacco L. & Ward M. (1996) "Temporal constraints on the McGurk effect", Perception and Psychophysics, 58, 351-362.

[Rad94a] Radeau M. (1994a) "Auditory-visual spatial interaction and modularity", Current Psychology of Cognition, 13, 1-51.

[Rad94b] Radeau M. (1994b) "Ventriloquism against audio-visual speech: Or, where Japanese speaking barn owls might help", Current Psychology of Cognition, 13, 124-140.

[Ros71] Rossi M. (1971) "L'intensité spécifique des voyelles", Phonetica, 24, 129-161.

[Sam97] Sams M., Surakka P., Helin P. & Kättö R. (1997) "Audiovisual fusions in Finnish syllables and words", AV Speech 1997, 101-104.

[Sek91] Sekiyama K. & Tohkura Y. (1991) "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility", Journal of the Acoustical Society of America, 90, 1797-1805.

[Sek93] Sekiyama K. & Tohkura Y. (1993) "Inter-language differences in the influence of visual cues in speech perception", Journal of Phonetics, 21, 427-444.

[Sum84] Summerfield Q. & McGrath M. (1984) "Detection and resolution of audio-visual incompatibility in the perception of vowels", Quarterly Journal of Experimental Psychology, 36A, 51-74.

[Vro92] Vroomen J. (1992), Hearing voices and seeing lips. Investigations in the psychology of lipreading. Doctoral Disseration, Tillburg University.

[Wal95] Walker S., Bruce V. & O'Malley C. (1995) "Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect", Perception and Psychophysics, 57, 1124-1133.

Remerciements

Notre recherche a été réalisée à l'aide des subventions du Ministère de la Recherche Scientifique de la Communauté Française de Belgique (Actions de Recherche Concertée; 96/01-203) et du Fonds National Belge de la Recherche Fondamentale Collective (8.4519.96). Nous remercions Willy Serniclaes pour ses précieux conseils et le C.A.V. de L'U.L.B qui a aidé au montage des stimuli.

Codage audiovisuel de la parole : Une étude préliminaire de la réduction du débit d'un vocodeur LPC par utilisation de paramètres vidéo

Elodie FOUCHER, Gang FENG & Laurent GIRIN

Institut de la Communication Parlée
INPG/ENSERG/Université Stendhal
B.P. 25, 38040 Grenoble CEDEX 09, France

foucher@icp.inpg.fr

Abstract

Today there exists numerous speech coding techniques which allow to transmit and stock efficiently the acoustic signals. But speech is both auditory and visual : audio and video information are complementary and the lip shape of the speaker can help the listeners to better understand the message. This paper aims to show the interest of video information in the speech coding domain in general and in terms of transmission rate in particular.

1. Introduction

La parole est un vecteur de communication bimodal : elle est à la fois auditive et visuelle. En effet, voir son interlocuteur, et plus particulièrement le mouvement de ses lèvres, permet, dans certaines situations, notamment bruitées, d'améliorer la compréhension de la parole. Cette bimodalité a déjà été exploitée dans les domaines de la reconnaissance et du débruitage des signaux de parole et les résultats obtenus ont été très prometteurs [Gir98a] [Rib95] [Adj97]. Il est donc très intéressant d'étudier un système de compression/codage du signal de parole exploitant la complémentarité entre le signal acoustique et l'image du locuteur. L'objectif de cette première étude est de prouver l'efficacité d'un système de codage audiovisuel en terme de débit de transmission.

L'apport de l'information visuelle est partiel : on ne peut évidemment pas reconstituer un signal de parole uniquement à partir de l'information vidéo. Mais on sait que voir son interlocuteur et notamment le mouvement de ses lèvres permet de mieux comprendre ce qu'il dit, donc améliore l'intelligibilité du message. En revanche, cette vision des lèvres n'apporte aucune information en terme de qualité du signal de parole émis. Ceci suggère que la modalité visuelle de la parole nous apporte des informations essentiellement relatives à l'allure spectrale globale du signal (enveloppe spectrale), synonyme d'intelligibilité de la parole, et non à la structure fine de celui-ci, essentiellement reliée à la source d'excitation et déterminant la qualité du signal de parole. Pour bien exploiter cette information visuelle, nous nous plaçons dans le cadre du codage à très bas débit qui privilégie l'objectif de l'intelligibilité du signal transmis par rapport à celui de la qualité. Ainsi, notre but est d'élaborer un système de codage audiovisuel améliorant les performances d'un codeur à bas débit. Pour réaliser cette étude, nous avons donc choisi de travailler sur un vocodeur à prédiction linéaire de référence à 2,4 kbits/s.

Dans cet article, nous proposons un vocodeur audiovisuel qui exploite l'information visuelle en vue d'une réduction du débit de transmission. Par rapport au vocodeur classique, le principe du codeur audiovisuel est, au cours de la phase de synthèse, d'estimer le filtre modélisant le conduit vocal à partir des paramètres acoustiques et visuels. Dans la section 2, nous allons décrire en détail la structure du codeur audiovisuel. Dans la section 3, nous aborderons l'estimation du filtre LPC avant de présenter dans les quatrième et cinquième sections le protocole d'expérimentation et les résultats obtenus.

2. Structure du codeur audiovisuel

Le principe classique d'analyse-synthèse trame par trame des codeurs LPC est conservé (figure 1). En amont, au cours de la phase de codage, une analyse LPC [Mar76] permet d'obtenir les coefficients du filtre 1/A(z) modélisant le conduit vocal. L'erreur de prédiction, ou résidu, est extraite par filtrage inverse du signal de parole à travers A(z). Cette erreur est ensuite modélisée classiquement par un bruit blanc ou un train d'impulsions, en fonction du voisement ou non de la trame étudiée. Les paramètres caractérisant le résidu, c'est-à-dire la fréquence fondamentale et l'énergie, sont transmis au décodeur.

Les paramètres audio utilisés dans cette étude sont les coefficients LAR (Log Area Ratio) qui sont plus adaptés à notre application et évitent tout problème de divergence des filtres. Ils se déduisent facilement des coefficients PARCOR de corrélation partielle par la relation suivante :

$$lar_i = \log\left(\frac{1+k_i}{1-k_i}\right)$$

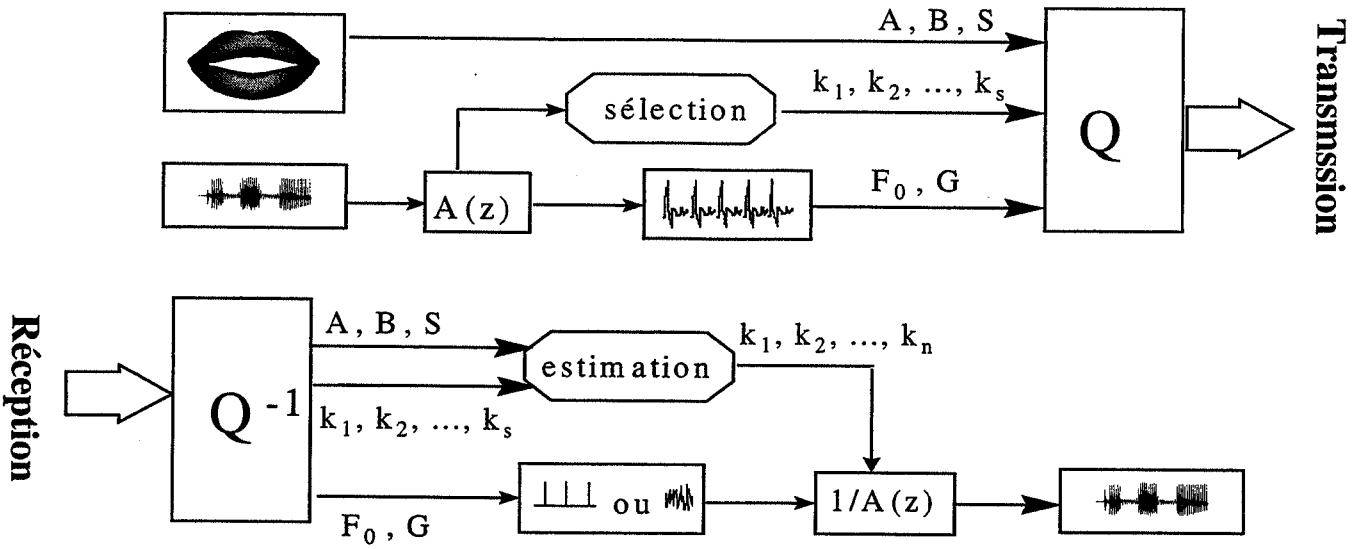


Figure 1 : Schéma de principe du vocodeur audiovisuel

Par ailleurs, le poste "visage-parole" de l'ICP [Lal91] nous permet d'extraire trois paramètres labiaux fondamentaux : l'étirement (A), la hauteur (B), et l'aire intérolabiale (S). Ces paramètres sont transmis, ainsi qu'une sélection des paramètres audio LPC du filtre 1/A(z), sélection destinée précisément à réduire le débit de transmission. Au cours du décodage, ce filtre est réestimé à partir des paramètres audio sélectionnés et des paramètres vidéo selon le processus décrit à la section suivante. Puis, le signal est synthétisé par filtrage de la source d'excitation, modélisée à partir des paramètres transmis, à travers le filtre 1/A(z) estimé à partir du jeu de coefficients audiovisuels.

3. Estimation du filtre 1/A(z) à partir des coefficients audiovisuels transmis

Le problème est ici d'estimer le filtre 1/A(z) à partir des coefficients transmis, c'est-à-dire un jeu de coefficients vidéo (V) et un jeu de coefficients audio (A1). La concaténation des vecteurs vidéo V et audio A1 en un vecteur audiovisuel est notée AV = [A1 V]. Nous devons, à partir de ce vecteur bimodal reçu, estimer le jeu de coefficients audio non transmis A2 qui, associé à A1, nous permettra de modéliser complètement le filtre de synthèse. La transmission du vecteur vidéo V demande un nombre de bits moins important que celle du vecteur audio A2. C'est donc dans un objectif de réduction du taux de transmission que cette première étude est réalisée.

Pour cela, nous utilisons un associateur linéaire, méthode qui s'est révélée simple et efficace dans plusieurs applications de traitement de la parole audiovisuelle, notamment le débruitage et la reconnaissance de parole [Gir98a] [Rob95]. C'est pourquoi on peut penser dans un premier temps qu'elle aura la même efficacité dans le domaine du codage de parole.

Son principe est le suivant : le vecteur A2 des coefficients audio non transmis est estimé par le produit du vecteur AV avec une matrice M. Cette matrice est déterminée lors d'une phase d'apprentissage par régression linéaire entre deux matrices contenant un large échantillon de vecteurs d'apprentissage des deux espaces "audio non transmis" et "audio transmis + vidéo". Cette méthode fonctionne suivant le principe de minimisation de l'erreur quadratique :

$$e = \| (AV)_{app} * M - (A2)_{app} \|.$$

Il est important de noter que les paramètres audio lar_i qui caractérisent le filtre de synthèse n'ont pas tous la même importance et ne sont donc pas codés avec le même nombre de bits. Ainsi, l'importance des coefficients dans la description de l'enveloppe spectrale décroît avec l'augmentation de leur indice. C'est pourquoi les premiers coefficients sont généralement codés avec plus de précision . C'est aussi pourquoi on retrouve ces premiers coefficients dans le vecteur A1 alors que A2 est constitué de la deuxième partie des coefficients lar_i dont l'importance et la précision de codage requise sont moindres.

Concernant le nombre de paramètres transmis, de nombreux cas de figures ont été testés en relation avec la qualité des signaux de parole obtenus en sortie du codeur. Comme la taille de A1 varie, le vecteur AV change et il existe donc autant d'associateurs que de cas de figures envisagés soit 10 associateurs différents (de 0 à 9 coefficients audio transmis).

4. Protocole d'expérimentation

Les corpus ont été choisis volontairement simples pour tester, dans un premier temps, la faisabilité et l'efficacité de notre système de codage audiovisuel. Cette étude étant totalement novatrice, il est important de travailler sur des signaux de parole simples pour voir quel type de difficultés on rencontre avant de tester notre codeur avec des signaux plus complexes.

C'est pourquoi, pour mettre en évidence l'efficacité du codeur audiovisuel, nous utilisons un corpus de 10 voyelles stationnaires [a e ɛ i ∅ œ o u y] et un corpus de transitions consonne-voyelle de la forme $V_1CV_2CV_1$ extraits des ensembles [a i u y] et [p t k b d g]. Pour chacun de ces corpus, nous disposons d'un corpus d'apprentissage, permettant de construire les associateurs audiovisuels par la méthode de régression linéaire vue au paragraphe précédent, et d'un corpus de test sur lequel les différentes expériences ont été réalisées.

Dans un premier temps, nous avons voulu montrer qu'il y a plus d'information en terme d'intelligibilité dans trois paramètres vidéo (A, B, S) que dans les trois premiers et principaux paramètres audio (lar_1, lar_2, lar_3) codés avec précision. Pour cela l'associateur est configuré pour passer d'un espace uniquement visuel à un espace uniquement audio. Un test de reconnaissance a été mis en place avec 10 auditeurs ayant entendu trois fois les 10 voyelles soit 30 stimuli.

Un deuxième type de test a été mis en place avec pour objectif mettre en évidence une diminution de débit de transmission tout en conservant une qualité de codage identique à celle du vocodeur classique. Cette fois-ci, un jeu de coefficients audiovisuels caractérisant la trame du signal de parole est transmis. On a fait varier le nombre de coefficients audio transmis en plus des paramètres vidéo jusqu'à obtenir un signal de qualité identique à celle d'un signal codé par un vocodeur classique. Un test auditif d'estimation de qualité entre différents signaux a été élaboré avec 10 auditeurs ayant entendu une fois le signal codé par vocodeur classique. Puis, ils ont désigné parmi les différents signaux de parole codés et reconstitués à partir de l'audiovisuel celui dont la qualité d'écoute était comparable à celle du vocodeur classique. On a alors relevé le nombre de coefficients audio transmis en plus des paramètres vidéo pour atteindre une qualité d'écoute identique à celle obtenue avec notre codeur de référence.

5. Résultats

La figure 2 présente les résultats obtenus au test d'intelligibilité sur les voyelles stationnaires avec la seule transmission des coefficients vidéo pour l'estimation du filtre LPC. Ainsi, à même nombre de paramètres transmis, trois coefficients vidéo permettent de restituer un signal intelligible alors que trois coefficients audio ne le permettent pas.

Un test d'intelligibilité a été réalisé sur le corpus des voyelles stationnaires (cf. figure 2) et sur le corpus des transitions consonne-voyelle et les conclusions sont identiques : à même nombre de paramètres transmis, on a plus d'information dans trois paramètres vidéo que dans trois paramètres audio.

L'apport visuel est donc évident mais reste partiel. En effet, un même jeu de paramètres labiaux peut décrire différentes voyelles, par exemple [e] et [i], ou différentes consonnes, par exemple [p] et [b], et on ne peut pas estimer deux filtres LPC différents avec un même jeu de paramètres visuels : il existe indiscutablement des ambiguïtés. C'est pour éviter ce type d'erreur qu'un certain nombre de coefficients audio est transmis avec les paramètres vidéo. On choisit de transmettre les premiers coefficients lar_i plus coûteux à coder mais essentiels pour caractériser la trame de parole et ainsi discriminer les " cas visuels identiques ".

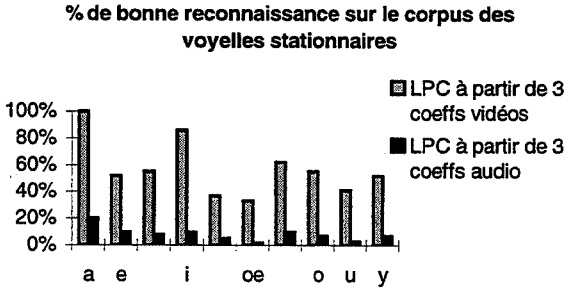


Figure 2

En ce qui concerne le deuxième test, qui consiste à conserver une qualité de codage identique à celle d'un vocodeur classique tout en réduisant le débit de transmission, on obtient des résultats apparaissant sur la figure 3 et relatif au corpus des voyelles stationnaires.

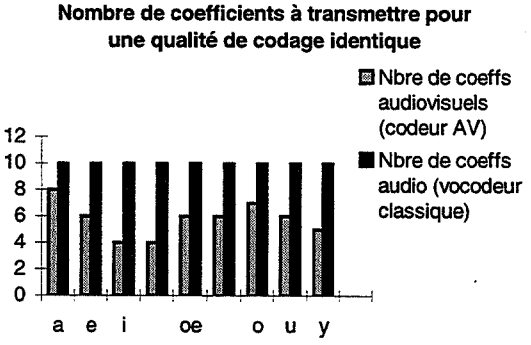


Figure 3

D'après les tests réalisés sur le corpus des transitions consonne-voyelle, nous pouvons conclure qu'il faut transmettre en moyenne 4 coefficients audio en plus des paramètres vidéo pour obtenir une qualité d'écoute du signal codé comparable à celle d'un signal codé par vocodeur classique.

En terme de débit de transmission, on peut quantifier la diminution du taux de transmission en fonction du nombre de coefficients audiovisuels transmis et du nombre de bits qui leur est alloué. En terme de débit de transmission, ces résultats apparaissent sur la figure 4. Par exemple, au lieu de transmettre 10 coefficients LPC dans le cas d'un vocodeur classique, on transmet 4 coefficients LPC et 2 paramètres vidéo dans le cas du codeur audiovisuel, soit 6 paramètres au lieu de 10. Dans ce cas, l'utilisation des paramètres audiovisuels permet de réduire de 40% le nombre de paramètres transmis. Ceci correspond à une diminution de 20 % du débit de transmission en fonction du nombre de bits alloués à chacun des paramètres transmis, tout en conservant une qualité de codage identique.

Evolution du débit (en kbit/s) du vocodeur audiovisuel en fonction du nombre de paramètres audiovisuels transmis

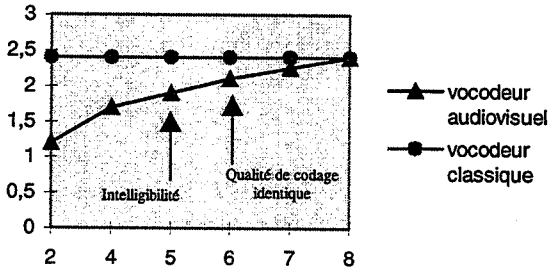


Figure 4

Notons que cette réduction de débit est obtenue malgré le fait que les coefficients audio transmis, en plus des paramètres visuels, sont les premiers lar_i qui demandent plus de précision de codage que les lar_i non transmis et estimés.

6. Discussion et perspectives

Nous avons montré, au cours de ces expériences, qu'il y a manifestement de l'information en terme d'intelligibilité dans les paramètres vidéo, et que la bimodalité du signal de parole peut être exploitée, dans un premier temps pour réduire le débit de transmission dans le cadre d'un codeur audiovisuel.

Ces résultats restent bien sûr très préliminaires mais ils confirment bien l'enjeu d'un codage audiovisuel et de l'utilisation conjointe de la bimodalité du signal de parole. L'aspect novateur de ce projet demande une certaine prudence dans notre travail ainsi que l'étude de cas très simples dans un premier temps pour tester la faisabilité et la crédibilité d'un tel système.

Les perspectives concernent les méthodes de construction des associateurs audiovisuels et notamment l'utilisation de réseaux de neurones permettant de réaliser des associations non-linéaires. Ils doivent être plus à même de mettre en rapport l'espace audio et l'espace vidéo, notamment dans les régions où la relation A/V linéaire est peu justifiée. Un premier test a été réalisé avec ce type d'associateur non-linéaire et les résultats obtenus sont encourageants [Gir98b].

Enfin, il est intéressant de se pencher sur la suppression de redondance entre certains paramètres audio et les paramètres vidéo pour permettre une efficacité maximale de notre système de codage/compression audiovisuel. La quantification vectorielle pourrait être une solution à ce problème [Ger92], en exploitant la complémentarité entre l'audio et vidéo pour la mettre au service d'une réduction de débit à qualité identique ou d'une augmentation de la qualité des signaux codés à débit de transmission équivalent, ceci est actuellement en cours d'étude.

L'intérêt de l'information vidéo dans le domaine du codage de parole nous apparaît donc très intéressant et de nombreuses pistes restent encore à explorer pour améliorer l'efficacité de notre système.

Bibliographie

[Ger92] GERSHO A., GRAY R.M. (1992), Vector Quantization and Signal Compression, Kluwer academic Publishers.

[Gir98a] GIRIN L., FENG G., SCHWARTZ JL., (1998), " Fusion of auditory and visual information for noisy speech enhancement : a preliminary study of vowel transitions", Proc. ICASSP'98, Seattle, USA.

[Gir98b] GIRIN L., VARIN L., FENG G., SCHWARTZ JL., (1998), "Débruitage audiovisuel de parole : apport d'une association vidéo-audio non-linéaire par des réseaux de neurones.", CORESA 1998.

[Lal91] LALLOUACHE M.T. (1991), " Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres ", Thèse de doctorat, Institut National Polytechnique de Grenoble.

[Mar76] MARKEL J.D., GRAY A.H. Jr (1976), Linear Prediction of Speech, Springer-Verlag, New-York.

[Rob95] ROBERT-RIBES J. (1995), Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception à la reconnaissance automatique des voyelles ", Thèse doctorale, INPG, Grenoble.

[Adj97] ADJOUDANI A. (1997), " Reconnaissance automatique de la parole audiovisuelle, Stratégies d'intégration et réalisation du LIPTRACK, labiomètre temps-réel ", Thèse doctorale, INPG, Grenoble.

Un modèle géométrique de lèvres 3D adaptable au locuteur

Lionel Revéret & Loïc Le Chevalier

Institut de la Communication Parlée
Université Stendhal / INPG-ENSERG
BP25X 38040 GRENOBLE CEDEX 9
Tél.: +33 4 76 82 41 28 - Fax: +33 4 76 82 43 35
e-mail : reveret@icp.inpg.fr - http://ophale.icp.inpg.fr/reveret.html

ABSTRACT

This papers presents a geometric 3D model of the lips. It is basically defined by a parametric surface. This surface interpolates a set of 30 control points in the XYZ space by polynomial equations. These control points are related to geometric features of the lips common to any human being. As such, this model can be adapted to most of the speakers' lip shapes. The purpose of our model is to provide a geometric representation of the lips for both synthesis and analysis of visual speech.

Firstly, we describe how the model is generated from the control points. Secondly, we present a graphical user interface which allows the model to be adjusted to the front and the profile views of a speaker's face. Finally, we present an automatic method for liptracking based on any 2D projection of the 3D model.

1. INTRODUCTION

Le gain d'intelligibilité apporté par la vision du visage du locuteur a été quantifié dans de nombreux travaux [Sum54, Nee56, Bin74, Sum79, Ben94]. Parmi les parties du visage entrant en compte dans la perception visuelle, les lèvres jouent un rôle prépondérant [Sum89, Ben96]. C'est pourquoi un nombre croissant d'études porte sur le traitement automatique de la parole visuelle, et en particulier sur la lecture labiale automatique : 15 références in [Sto96], [Rog96, Jou96, Bas97] par exemple. Il existe plusieurs approches à la "labiométrie" automatique. À l'ICP, ont été en particulier développées des méthodes basées sur un traitement noir et blanc [Abr86, Lal88] et couleur de lèvres maquillées [Lal91], ainsi que, pour des images de lèvres non maquillées, un traitement stochastique au niveau pixel [Rev97a] et une application en modèle déformable d'un modèle 3D de lèvres [Rev97b].

Jusqu'à présent, notre modélisation de lèvres par contours actifs n'était développée que sur un seul locuteur. Le modèle utilisé était issu de la synthèse de lèvres 3D parlantes [Gui96]. Afin de généraliser cette approche dans une perspective de labiométrie multilocuteur, nous avons développé un deuxième prototype de modèle géométrique de lèvres 3D susceptible d'être adapté à différentes formes de lèvres.

Dans un contexte de communication homme-machine, la modélisation labiale s'applique à la fois à la synthèse de visages parlants et à la lecture labiale automatique. C'est pourquoi le modèle géométrique de lèvres présenté dans cet article vise à être appliqué simultanément à l'analyse et à la synthèse de lèvres parlantes.

2. MODÉLISATION GÉOMÉTRIQUE

Nos modèles de lèvres sont définis par une surface paramétrique 3D. La surface du nouveau modèle est construite par interpolation entre des points de contrôle.

2.1 Points de contrôle

Les points de contrôle retenus sont communs à n'importe quelle forme de lèvres humaines. Ils sont répartis le long de 3 contours : externe, interne, et médian. 10 points suffisent à caractériser chaque contour. Au sein d'un contour, chaque point correspond à un lieu géométrique : un point marque chaque commissure des lèvres, trois points l'arc de cupidon, un point la limite inférieure du contour, les quatre autres sont intermédiaires entre ces zones : voir figure 1.

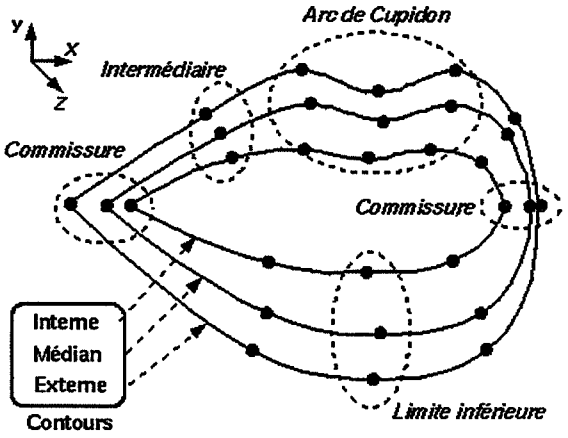


Figure1: Les 30 points de contrôle.

Par ces 30 points dans l'espace XYZ, le modèle de base est contrôlé a priori par 90 paramètres de

commande. La définition du contour interne n'est pas anatomique : elle correspond à la limite visible de l'intérieur de la cavité buccale dans une vue de face.

2.2 Surface d'interpolation

La surface est engendrée par le parcours d'une courbe 3D circulant du contour externe vers le contour interne en passant par le contour médian. Cette courbe polynômiale, continue par morceaux, passe également par dix points de contrôle, calculés par interpolation polynômiale entre les points correspondants des trois contours initiaux.

Les courbes polynômiales 3D proviennent de la combinaison de courbes dans le plan coronal XY, et dans le plan axial XZ. Les courbes dans le plan XY sont une généralisation des courbes du modèle 3D précedent [Gui96], lui même inspiré du modèle proposé par Lindblom [Lin71]. Les courbes dans le plan XZ sont elles aussi une généralisation des courbes utilisées dans le modèle précédent [Adj94] : voir figure 2. La surface paramétrique d'interpolation obtenue finalement est présentée à la figure 3.

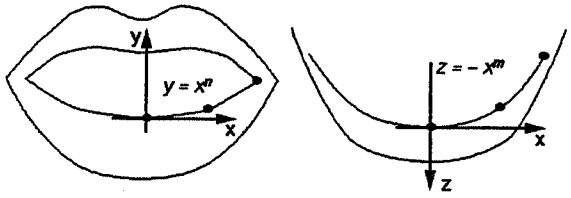


Figure 2: Courbes polynômiales dans les plans XY et XZ.

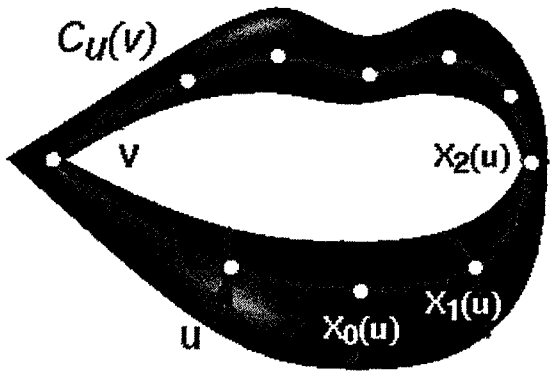


Figure 3: La surface paramétrique en rendu « Gouraud shading ».

3. Reconstruction 3D Interactive

3.1 Interface de reconstruction

Chaque locuteur est enregistré de face et de profil, les deux vues étant mixées sur une même image. Une interface graphique permet une reconstruction 3D complète par mise en correspondance des deux vues des lèvres, connaissant la position dans l'espace de chaque caméra par rapport au locuteur. En ajustant les points de contrôle, le modèle est mis à jour et projeté sur les vues originales.

La figure 4 présente un exemple du modèle, projeté sur des lèvres vues de face et de profil, dans lequel un point de contrôle a été délibérément déplacé.

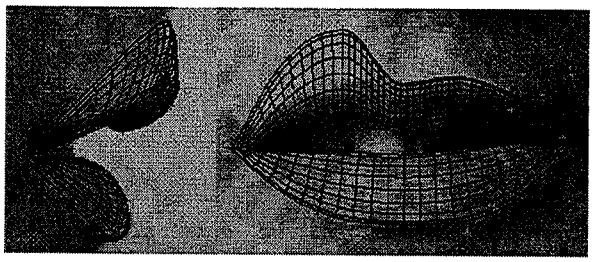


Figure 4: Placement interactif des points de contrôle, rendu du modèle en « fil de fer ».

3.2 Contraintes géométriques

Le placement de certains points de contrôle est ambigu : le profil non filmé est intégralement caché par l'autre profil ; le contour médian est mal défini sur la vue de face. Des règles de géométrie ont donc été imposées afin de réduire les ambiguïtés et forcer le positionnement de certains points en fonction d'autres.

A été laissé à l'utilisateur l'ajustement des points ne présentant que peu d'ambiguïtés : positionnement en XY des points des contours interne et externe ; en Z des points de contrôle sur le contour externe ; en Z des deux autres points du bord visible de l'arc de Cupidon ; en Z des limites inférieures des contours interne et médian.

Les autres paramètres sont déduits par des règles géométriques : positionnement en XY du contour médian au milieu des contours interne et externe ; symétrisation des ordonnées Z des deux profils ; l'ordonnée Z des points intermédiaires, interne et médian, est calculée par apposition d'un coefficient de courbure déduit du contour externe.

3.3 Paramétrisation

Afin de diminuer les degrés de libertés de notre modèle, nous avons appliqué une analyse en composantes principales (ACP) sur les points de contrôle du modèle placés manuellement sur des lèvres représentatives du français incluant les extrêmes de fermeture / ouverture et d'arrondissment / étirement. Le modèle a été appliqué à 23 images face et profil correspondant aux visèmes identifiés par [Ben92] sur ce corpus.

Les points de contrôle ont été recadrés par rapport à un point fixe de la tête. Une ACP a été effectuée sur les 90 coordonnées caractéristiques de chaque visème. Le premier vecteur propre correspond à un geste d'étirement / protrusion et compte pour 72% de la

variance, le second à un geste de fermeture / ouverture (18%). Le troisième ne compte que pour 3% de la variance. Voir figure 5.

Ce modèle, en cours de construction, n'a pour le moment été entraîné que sur un seul locuteur, et les résultats ci- dessus restent à vérifier sur un plus grand groupe de sujets. L'analyse multilocuteur permettra aussi de réaliser l'objectif initial de notre étude, à savoir une paramétrisation inter-locuteur des formes labiales. Néanmoins, pour illustrer d'ores et déjà l'adaptabilité de notre modèle à différentes conformations labiales, la figure 6 montre comment notre modèle s'ajuste aux lèvres de trois sujets différents prononçant la voyelle /a/.

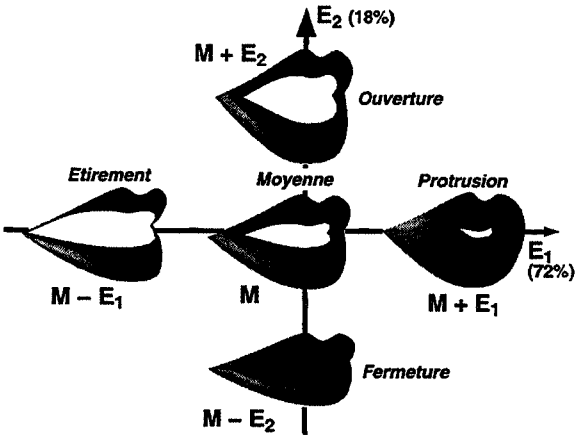


Figure 5: Variation selon les deux premiers vecteurs propres d'une ACP sur 23 images d'un locuteur.

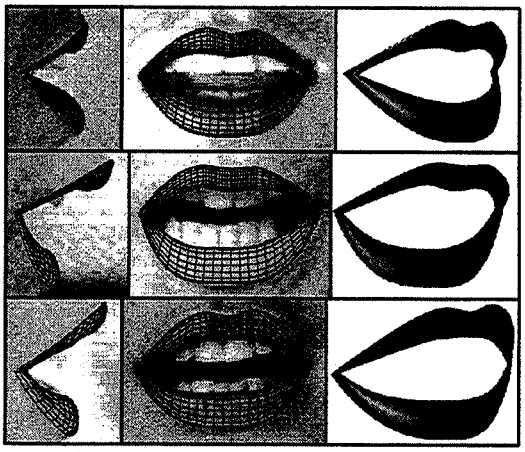


Figure 6: Applications du modèle à différents locuteurs.

En attendant, ce premier résultat fait émerger une paramétrisation articulatoire intéressante de notre modèle pour la synthèse. Il est en effet possible de créer une nouvelle forme du modèle par combinaison linéaire des deux vecteurs propres (orthogonaux par définition) :

$$\mathbf{X} = \mathbf{M} + \alpha\mathbf{E}_1 + \beta\mathbf{E}_2$$

où **M** est la moyenne des positions XYZ des points de contrôle pour les 23 visèmes, $\mathbf{E}_1$ et $\mathbf{E}_2$ les deux vecteurs propres, α un coefficient d'étirement / protrusion, β un coefficient d'ouverture / fermeture et **X** la position des 30 points de contrôle pour le calcul du modèle complet.

4. Reconstruction 3D Automatique

Dans cette partie nous présentons un algorithme de suivi automatique du modèle 3D sur une séquence video du locuteur.

4.1 Suivi de contours en 3D

L'utilisation de modèles 3D en suivi de contour présente, par rapport aux méthodes 2D classiques [Yui92, Lue97], l'avantage de pouvoir prendre en compte la rotation de la tête du locuteur et les différences d'angle de vues [Bas98, Rev97b]. Dans notre cas, les paramètres du modèle sont itérativement ajustés jusqu'à ce que les contours apparents du modèle se situe dans une zone de gradient maximal, *quel que soit l'angle de vue.*

A chaque image, un prétraitement de la couleur est appliqué pour isoler les lèvres. Un calcul de gradient estime la zone de leurs contours apparents. En fonction de l'angle de vue, le modèle 3D est projeté selon un modèle optique de caméra. L'adéquation du modèle à la position réelle des lèvres est évaluée par la mesure du gradient le long des contours apparents du modèle dans cette vue (figure 7).

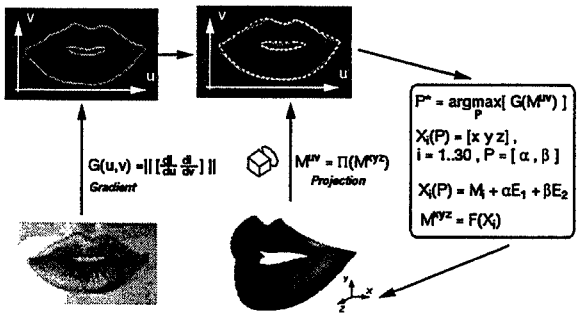


Figure 7: Principe du suivi de contour 3D

Malgré les règles de géométrie exposées ci-dessus, les 90 paramètres initiaux restent en nombre trop élevé pour permettre une convergence efficace du modèle. Les deux premiers vecteurs propres ci-dessus ont été utilisés comme modes de déformation du modèle. Le suivi de contour consiste donc à retrouver à partir de l'image la valeur des deux paramètres de protrusion et d'ouverture évoqués dans la section précédente. Bien entendu, ce suivi ne peut s'appliquer qu'à notre premier locuteur dans l'état actuel de notre modèle.

Cette paramétrisation « articulatoire » du mouvement des lèvres fait l'hypothèse d'une déformation recadrée par rapport à la tête du locuteur. Tout mouvement rigide de la tête doit donc être rigoureusement éliminé.

La plupart des travaux en suivi de contour intègrent dans les modes de déformation du modèle une paramétrisation de ce mouvement rigide extralinguistique. Nous pensons au contraire que le contrôle des mouvements de tête doit être pris en compte par un système indépendant [Rev97b].

4.2 Exemples

La figure 8 montre trois images issues d'une séquence /azyzaz/ de notre locuteur de référence, maquillé, tête maintenue par un casque solidaire de la chaise, vu de face et profil.

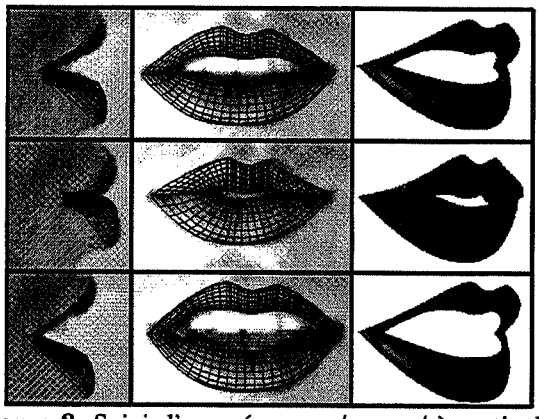


Figure 8: Suivi d'une séquence /azuzaz/ à partir des vues de face et profil.

Dans la séquence présentée, l'ajustement du modèle combine une optimisation simultanée de face et de profil. En ne retenant que la vue de face les résultats sont légèrement moins bons (retard de la fermeture). Un suivi sur la vue de profil seule ne parvient pas à converger correctement sur la position des lèvres.

5. CONCLUSION

Nous avons présenté dans cet article un modèle géométrique 3D de lèvres pour l'analyse / synthèse. Ce modèle est adaptable à la fois aux déformations d'un locuteur et aux conformations de différents locuteurs.

Les résultats de l'ACP effectuée sur un seul locuteur sont prometteuses pour le contrôle de lèvres parlantes (synthèse) comme pour la labiométrie automatique (analyse) à très petit nombre de degrés de liberté. Notre méthodologie ouvre une perspective originale vers une paramétrisation multi-locuteur en suivi de contours labiaux ainsi qu'une possible paramétrisation de la morphologie labiale caractéristique du locuteur.

BIBLIOGRAPHIE

[Abr86] C. Abry, L.J. Boë (1986), "Laws for lips", Speech Communication, 5, 97-104.

[Adj94] A. Adjoudani, T. Guiard-Marigny, B. Le Goff, C. Benoît (1994), "Un modèle 3D de lèvres parlantes", Actes des XX° JEPs, 143-146.

[Bas98] S. Basu, N. Oliver, A. Pentland (1998), "3D modeling and tracking of human lip motions", ICCV'98.

[Ben92] C. Benoît, M.T. Lallouache, C. Abry (1992), "A set of French visemes for visual speech synthesis", Talking Machines : Theories, Models and Designs, 485-504, Elsevier.

[Ben94] C. Benoît, T. Mohamadi, S. Kandel (1994), "Effects of phonetic context on audio-visual intelligibility of French", J. Speech & Hearing Res., 37, 1195-1203.

[Ben96] C. Benoît, T. Guiard-Marigny, B. Le Goff, A. Adjoudani (1996), "Which components of the face do humans and machines best speechread ?", Speechreading by Humans and Machines, Springer-Verlag, 351-372.

[Bin74] C.A. Binnie, A.A. Montgomery, P.L. Jackson (1974), "Auditory and visual contributions to the perception of consonants", J. Speech & Hearing Research, 17, 619-630.

[Gui96] T. Guiard-Marigny (1996), "Modélisation tridimensionnelle des articulateurs de la parole", Thèse de Doctorat de l'INPG.

[Jou96] P. Jourlin (1996), "Asynchronie dans les systèmes de reconnaissance de la parole basés sur les HMM", Actes des XXI° JEPs, 351-354.

[Lal88] M.T. Lallouache, C. Worley (1991), "Saisie, édition et traitement d'images et de signaux articulatoire lèvres et mâchoire", Actes des XVI° JEPs, 113-115.

[Lal91] M.T. Lallouache (1991), "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", Thèse de Doctorat de l'INPG.

[Lin71] B. Lindblom, J. Sundberg (1971), "Acoustical consequences of lip, tongue, jaw, and larynx movement", J. of the Acoustical Society of America, 59, 1166.

[Nee56] K.K. Neely (1956), "Effect of visual factors on the intelligibility of speech", J. Acoustical Society of America, 28, 1275-1277.

[Rev97a] L. Revéret (1997), "From raw images of the lips to articulatory parameters : a viseme-based prediction", Eurospeech'97, 4, 2011-2014.

[Rev97b] L. Revéret, F. Garcia, C. Benoit, E. Vatikiotis-Bateson (1997), "An hybrid approach to orientation-free liptracking", AVSP'97, 117-120.

[Rog96] A. Rogozan, P. Deléglise, M. Alissali (1996), « Intégration asynchrone des informations auditives et visuelles dans un système de reconnaissance de la parole », Actes des XXI° JEPs, 359-362.

[Sto96] D. Stork, M. Hennecke Eds. (1996), « Speechreading by Humans and Machines », NATO-ASI series Vol. 150, Springer-Verlag.

[Sum54] W.H. Sumby, I. Pollack (1954), "Visual contribution to speech intelligibility in noise", J. Acoustical Society of America, 26, 212-215.

[Sum79] Q. Summerfield (1979), "Use of visual information for phonetic perception", Phonetica, 36, 314-331.

[Sum89] Q. Summerfield, A. MacLeod, M. McGrath, M. Brooke (1989), « Lips, teeth, and the benefits of lipreading », Handbook of Research on Face Processing, Elsevier, 223-233.

[Yui92] A. L. Yuille, P.W. Hallinan, D.S. Cohen (1992), « Feature extraction from faces using deformable templates », Int. J. Computer Vision, 8:2, 99-111.

Apports d'un apprentissage discriminant du visuel à la reconnaissance audiovisuelle de la parole

Alexandrina ROGOZAN et Paul DELÉGLISE

Laboratoire d'Informatique de l'Université du Maine
Université du Maine - 72085 Le Mans Cedex 9
Tél. : +00 2 43 83 38 64 - Fax : +00 2 43 83 38 68
e-mail : afoucaul@lium.univ-lemans.fr

Abstract

In this paper we present an improved version of our audio-visual speech recognition system developed according to a hybrid (DI + SI) integration model. For the purely-visual recogniser belonging to the SI component, our approach consists in using a viseme set adapted to the recognition task and neural networks trained to discriminate between visemes. The system performance is tested on a connected-letter speech recognition task under varying noise-level conditions and then compared to those obtained with the precedent-version systems.

1. Introduction

De nombreuses études ont montré l'importance des informations visuelles pour la perception de la parole [Rob96]. Ces résultats ont été aussi confirmés dans le cas des systèmes automatiques de reconnaissance audiovisuelle [Adj96 ; And97 ; Jou97 ; Tei97]. Les expériences que nous avons réalisées [Rog97b] sur ces systèmes ont montré une supériorité des systèmes fondés sur le modèle d'intégration hybride (ID+IS) par rapport au modèle d'intégration directe (ID) et au modèle d'intégration séparée (IS).

Nous proposons dans cet article une amélioration du système hybride par la définition et l'utilisation d'un ensemble de visèmes adapté à la tâche de reconnaissance. De plus, pour renforcer l'apport de ces visèmes nous proposons leur utilisation dans le cadre d'un apprentissage discriminant à base de réseau neuro-mimétique. La fusion du modèle DI avec les informations labiales se fait en comparant la solution du réseau neuro-mimétique avec les N Meilleures solutions du système markovien représentant le modèle ID. Le critère de décison fait intervenir une mesure de fiabilité sur les résultats des deux sous systèmes de reconnaissance.

Dans les sections suivantes nous présentons donc la détermination des visèmes suivie de l'architecture du système audiovisuel de reconnaissance et pour finir les résultats obtenus avec ce système et leur analyse.

2. Détermination des visèmes

L'utilisation des phonèmes comme unités de décision dans un système purement visuel de reconnaissance de la parole n'est pas souhaitable. En effet, des phonèmes différents peuvent être visuellement similaires, *e. g.* les phonèmes {p, b, m}. De ce fait, d'autre unités ont été proposées pour étiqueter le continuum visuel : les visèmes. Ils représentent l'ensemble des unités distinctives et indissociables qui décrivent le signal de parole sur le plan visuel [Ben92]. Les visèmes peuvent aussi être déterminés par regroupement des phonèmes en classes selon leur similarité visuelle [Gol96].

Cependant, la classification des phonèmes en visèmes ne peut pas être complètement déterminée *a priori* car la réalisation visuelle d'un phonème est dépendante des modes d'articulation des locuteurs [Kri96]. De ce fait, le nombre de classes visémiques et leurs constituants est sujet à variation selon le locuteur. Une méthode automatique pour déterminer les visèmes les plus adaptés au locuteur pour une tâche de reconnaissance est donc souhaitable ; il faut donc utiliser une méthode de classification automatique sur le corpus d'apprentissage de notre base audiovisuelle.

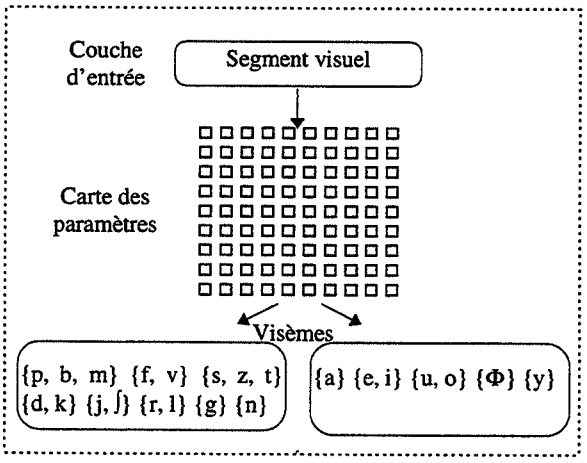


Figure 1 : Regroupement des phonèmes en visèmes

Dans un premier temps cela nécessite de trouver une représentation des éléments à classer. Pour cela une segmentation semi-automatique est effectuée sur le signal acoustique à partir de la transcription phonétique des phrases. Puis compte tenu des asynchronismes [Abr91] entre les effets visuels et les effets acoustiques d'un phonème, nous avons décidé de ne pas projeter les frontières acoustiques sur le signal visuel. Nous avons seulement reporté le milieu des segments acoustiques sur ce signal pour ancrer des segment visuels de taille fixe suffisamment grande. Cette dernière a été choisie à 140 ms pour

que, d'une part, la réalisation visuelle associée au phonème soit obligatoirement présente dans le segment et que, d'autre part, une partie des transitions vers les phonèmes voisins soit incorporée dans le segment. Ainsi, la classification tiendra compte des phénomènes de coarticulation [Ben94] qui influent aussi sur les processus de reconnaissance.

Dans un deuxième temps nous devons classer ces segments sans introduire trop d'*a priori* dans la classification. Pour cette tâche nous devons donc utiliser un algorithme non supervisé. Parmi les nombreux algorithmes existants nous avons choisi celui de Kohonen [Koh88] sur les cartes auto-organisatrices (cf. Figure1), ce type d'algorithme ayant déjà fait ces preuves dans les problèmes de regroupement de données multidimensionnelles [Dem97]. Pour simplifier le processus de classification, nous traitons de manière séparée les consonnes et les voyelles en posant comme *a priori* qu'un visème ne pouvait contenir à la fois une voyelle et une consonne.

Cette étape de classification fournit treize visèmes dont les constituants sont présentés dans la Figure 1. On peut constater que ces visèmes sont cohérents avec ceux trouvés par d'autres chercheurs [Gol96 ; Kri96] ce qui est une première validation de la méthode, la deuxième provenant des bons résultats des méthodes de reconnaissance utilisant ces visèmes [Rog97a].

3. Architecture hybride markovienne connexioniste pour la reconnaissance audiovisuelle de la parole

A la suite des études présentés dans [Rog97b] nous avons choisi d'utiliser un système hybride (ID+IS) où l'intégration des informations acoustiques et visuelles s'effectue à la fois au niveau pré-catégorique (ID) et aussi au niveau post-catégorique (IS). Dans ce qui suit, nous allons donc définir le système ID, la partie visuelle du système IS, ainsi que leur intégration.

3.1 Système ID

Le système ID utilise des modèles de Markov cachés continus (MMCC) phonétiques, dont l'observation bimodale d'entrée est obtenue par concaténation pondérée des paramètres visuels et acoustiques. Pour permettre son intégration dans le système IS, la sortie de ce sous-système n'est pas seulement constituée de l'énoncé le plus probable mais de la suite des N énoncés les plus probables et phonétiquement différents. Chaque énoncé est constitué de la suite des phonèmes et de leur position.[Ali96].

3.2 Composante visuelle du système IS

Pour le système IS nous avons décidé d'utiliser les informations visuelles de manière discriminante puisqu'il s'agit, à cette étape seulement, de choisir parmi les N meilleurs énoncés proposés par le système ID. Pour cela, l'utilisation de réseaux neuro-mimétiques (RN) entraînés pour la reconnaissance des visèmes a été choisie.

Afin de traiter la dimension temporelle des informations visuelles il est possible d'utiliser soit des réseaux multicouches à retard, soit des réseaux récurrents. Des expériences présentées dans [Rog98] montrent que pour ce contexte les réseaux récurrents [Jor86] ont de meilleurs résultats ; c'est donc ces derniers qui ont été préférés pour nos expériences.

Pour constituer l'ensemble d'apprentissage, les segments visuels sont extraits du corpus de la même manière que dans la Section 2. L'apprentissage a été réalisé par l'utilisation d'une version adaptée de l'algorithme de rétropropagation de l'erreur. Le choix de paramètres a été réglé sur l'ensemble de validation avant d'utiliser le RN pour la reconnaissance.

3.3 Intégration du ID avec la composante visuelle du IS

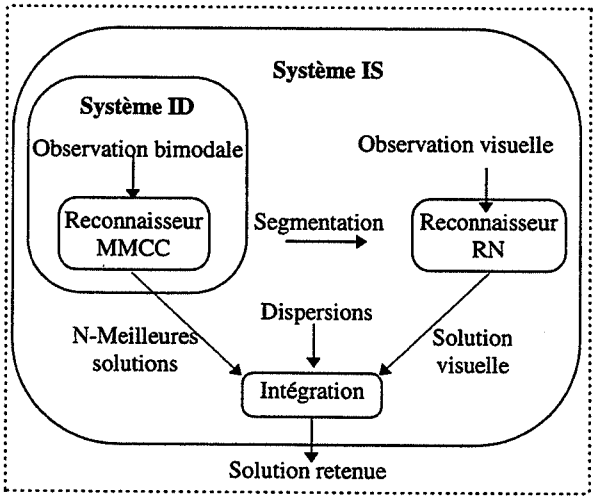


Figure 2 :Architecture hybride markovienne connexioniste pour la reconnaissance audiovisuelle de la parole

L'intégration du système ID avec la composante visuelle du système IS est schématisée en Figure 2. Nous avons choisit le système ID comme référence parce qu'il est le plus fiable pour cette tâche de reconnaissance.

Dans un premier temps la segmentation correspondante à la meilleure solution proposée par le système ID est utilisée pour extraire du corpus de test un ensemble de segments visuels de taille fixe qui seront ensuite étiquetés par le réseau récurrent. Pour réaliser cette extraction, nous avons utilisé la méthode choisie pour la détermination des visèmes (projection des milieux des segments ...).

Dans un deuxième temps une mesure de fiabilité est effectuée pour le système ID et pour la partie visuelle du système IS. Cette mesure est définie comme la dispersion des résultats. Cette dernière est calculée à partir des différences des scores des quatre meilleures solutions [Adj96 ; Rog97b]. Ainsi un critère de fiabilité pour un processus de reconnaissance peut être défini : nous considérons que ce

processus est fiable si la valeur de sa dispersion sur la phrase est au-dessus d'un certain seuil qui est fixé à la valeur moyenne de la dispersion sur le corpus de test. Nous utilisons donc ce critère pour la décision finale :

- Si la reconnaissance du système ID est fiable ou si la reconnaissance purement visuelle n'est pas fiable, la solution retenue est le meilleur énoncé proposé par le système ID ;
- sinon la composante visuelle du système IS choisit parmi les N plus probable énoncés fournis par ID celui qui est le plus semblable à la solution visuelle.

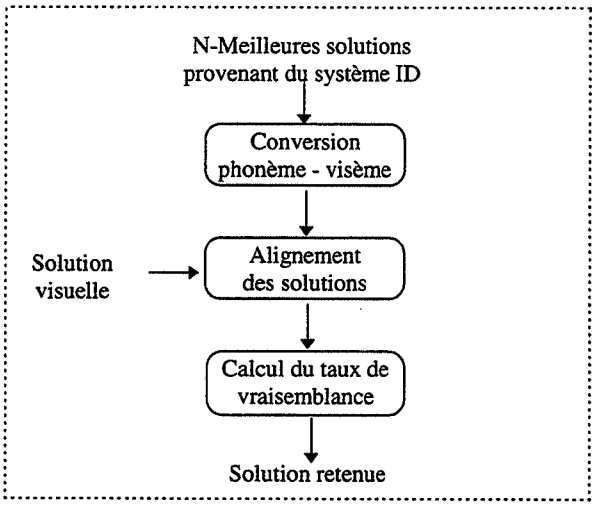


Figure 3 : Discrimination visuelle des N meilleurs énoncés provenant du système ID

Le processus de discrimination visuelle des N plus probables énoncés provenant du sous-système ID est présenté dans la Figure 3. Ces N énoncés sont convertis en séquences de visèmes et alignés par un algorithme de distance d'édition sur la suite fournit par le sous-système visuel. Cet alignement fournit un taux de vraisemblance pour chacun des énoncés ; ce taux indique la distance entre l'énoncé et la solution du système visuelle. L'énoncé retenu correspond à celui qui est le plus proche de la solution visuelle au sens de ce critère.

4. Expériences et résultats

Nous avons étudié les performances du système hybride (ID+IS) à composante visuelle discriminante comparativement au système acoustique et au système visuel sur la tâche suivante : reconnaissance de séquences de quatre lettres prononcées de façon continue par un seul locuteur, extraites d'un corpus audiovisuel.

Réalisé à l'ICP de Grenoble, ce corpus est composé de 200 phrases et représente l'enregistrement synchrone des informations acoustiques et des informations de suivi des lèvres. Les paramètres visuels que nous avons utilisés pour ces expériences correspond à l'étirement, la séparation et l'aire intero-labiale. Pour étudier les performances des systèmes en milieu bruité, nous avons ajouté au signal acoustique un « bruit de foule ». Le niveau de bruit correspond à un rapport du signal sur bruit (RSB) de 10 dB, 0 dB et -10 dB.

L'apprentissage des MMCC a été réalisé sur deux tiers du corpus. Seulement un tiers a été utilisé pour l'apprentissage des RN, puisque un autre tiers devait servir pour la validation. Les différents systèmes ont été testés sur le tiers restant du corpus. Dans la Table 1 nous présentons les résultats exprimés en terme du taux de lettres correctement reconnues moins le taux d'insertions pour les systèmes audiovisuels et acoustiques dans diverses conditions de bruit. Les résultats des systèmes purement visuels sont exprimés en taux de visèmes, la tâche de reconnaissance étant effectuée dans ce cas sans lexique.

Table 1 : Résultats des tests

Lettres / Visèmes %	sans bruit	10 dB	0 dB	-10 dB
Systèmes unimodaux				
Acoustique (A)	90.8	85.5	67.9	-44.3
Visuel (V)	52.8	51.7	49.3	40.0
Systèmes bimodaux				
(ID)	95.4	88.3	75.0	37.6
(ID + IS markovienne)	95.4	89.0	77.8	37.6
(ID + IS connexioniste)	95.7	89.8	75.7	38.0

Les performances du système hybride utilisant un apprentissage discriminant pour le visuel avec différentes RSB sont présentées sous la ligne (ID + IS connexioniste) dans la Table 1. En vue de l'évaluation de l'apport de l'apprentissage discriminant, les résultats d'un système hybride utilisant pour la partie visuelle de IS un apprentissage markovien [Rog97b] ont été incluses dans le tableau des performances sous la ligne (ID + IS markovienne).

5. Discussion et conclusion

On peut remarquer que, pour tous les RSB testés, le résultat de la fusion est supérieur à ceux des deux modalités prises séparément en tenant compte du fait que les résultats du système purement visuel est exprimé en taux de visèmes et non pas en taux de lettres comme pour les autres systèmes

De plus, ces performances confirment encore les résultats précédents sur la supériorité des systèmes hybrides (ID+IS) sur le système ID. La comparaison des deux dernières lignes montre un léger avantage pour l'apprentissage discriminant. En faveur du système hybride utilisant un apprentissage discriminant on peut noter que :

- l'apprentissage n'a pu être effectué seulement que sur un tiers du corpus, alors que le double a été utilisé pour l'apprentissage markovien ;

- la meilleure solution du système ID est favorisée dans le processus de fusion puisque le reconnaisseur visuel n'utilise, à l'heure actuelle, que seulement cette segmentation ;

- la construction de la solution visuelle n'utilise pas le lexique, contrairement au reconnaisseur markovien.

Dans cet article nous avons montré que l'on peut tirer parti de l'utilisation d'un apprentissage discriminant des informations visuelles pour la reconnaissance audiovisuelle de la parole. En outre, la méthode que nous avons proposée permet la fusion de deux approches différentes : connexioniste et stochastique appliquées à la reconnaissance acoustique et visuelle. Cependant, les résultats obtenus n'étant pas toujours statistiquement importants, les différents systèmes devront être testés sur un corpus plus conséquent.

Bibliographie

[Abr91] Abry C. and Lallouache T. (1991), « Audibility and Stability of Articulatory Movements: Deciphering Two Experiments on Anticipatory Rounding in French », in Proc. of XIIth International Congress of Phonetic Sciences, Aix -en - Provence, France, 1991, pp. 220-225.

[Adj96] Adjoudani A. and Benoît C. (1996), « On the Integration of Auditory and Visual Parameters in an HMM-based ASR », in: D. Stork and M. Hennecke, eds., Speechreading by Humans and Machines, (NATO ASI Series. Series F: Computer and Systems Science, Vol. 150), pp. 461-473.

[Ali96] M. Alissali, P. Deléglise and A. Rogozan « Asynchronous Integration of Visual Information in an Automatic Speech Recognition System », in Proceedings of *ICSLP'96*, Philadelphy, U.S.A., October, 1996.

[And97] André-Obrecht R., Jacob B. and Parlangeau N. (1997), « Audio-Visual Speech Recognition and Segmental Master-Slave HMM », in: C. Benoît and R. Campbell, eds., Proc. of the ESCA/ESCOP workshop on Audio-Visual Speech Processing, Rhodes, Greece, 26-27 September 1997, pp. 49-52.

[Ben92] Benoît C., Lallouache T., Mohamadi T. and Abry C. (1992), « A Set of French Visemes for Visual Speech Synthesis », in: G. Bailly and C. Benoît, eds., Talking machines: Theories, Models, and Designs, (Elsevier Science Publishers, Amsterdam), pp. 485-504.

[Ben94] Benoît C., Mohamadi T. and Kandel S. (1994), « Effects of Phonetic Context on Audio-Visual Intelligibility of French », Journal of Speech and Hearing Research, vol. 37, pp . 1195-1203.

[Dem97] Demartines P and Hérault J (1997), « Curvilinear Component Analysssis / A Sekf-Organising Neural Network for Nonlinear Mapping of Datasets », IEEE Transactions on Neural Networks, vol 8 n 1 Jan 97, pp 148-164

[Gol96] Goldschen A., Garcia O. and Petajan E. (1996), « Rationale for Phoneme-Viseme Mapping and Feature Selection in Visual Speech Recognition », in Speechreading by Humans and Machines, pp. 505-509.

[Jor86] Jordan M. I. (1986), « Attractor Dynamics and Parallelism in a Connectionist Sequential Machine », Proc. of the Eighth Annual Conference of the Cognitive Science Society, pp. 531-546.

[Jou97] Jourlin P. (1997), « Word-Dependent Acoustic-Labial Weights in HMM-based Speech Recognition », Proc. of the ESCA/ESCOP workshop on Audio-Visual Speech Processing, Rhodes, Greece, 26-27 September 1997, ed. by C. Benoît and R. Campbell, pp. 69-72.

[Koh88] Kohonen T. (1988), Self-Organization and Associative Memory, Springer-Verlag.

[Kri96] Kricos P. B. (1996), « Differences in Visual Intelligibility Across Talkers », in: D. Stork and M. Hennecke, eds., Speechreading by Humans and Machines (NATO ASI Series. Series F: Computer and Systems Science, Vol. 150), pp. 43-55.

[Rob96] Robert-Ribes J., Piquemal M., Schwartz J.L. and Escudier P. (1996), « Exploiting Sensor Fusion Architectures and Stimuli Complementarily in AV Speech Recognition », in: D. Stork and M. Hennecke, eds., Speechreading by Humans and Machines (NATO ASI Series. Series F: Computer and Systems Science, Vol. 150), pp. 193-210.

[Rog97a] Rogozan A. and Deléglise P. (1997), « Continuous Visual Speech Recognition Using Geometric Lip-Shape Models and Neural Networks », in Proc. of the 5th Conference on Speech Communication and Technology, Rhodes, Greece, 26-27 September 1997, pp. 1999-2003.

[Rog97b] Rogozan A., Deléglise P. and Alissali M. (1997), « Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition », in: C. Benoît and R. Campbell, eds., Proc. of the ESCA/ESCOP workshop on Audio-Visual Speech Processing, Rhodes, Greece, 26-27 September 1997, pp. 61-65.

[Rog98] Rogozan A. and Deléglise P. (1998), « Hybrid Hidden Markov Models / Neural Network Models for Speechreading », accepted to European Symposium on Artificial Neural Networks, Bruges, Belgium, 22-24 April 1998.

[Tei97] Teissier P., Robert-Ribes J., Schwartz J.L. and Guérin-Dugué A. (1997), « Models for Audio-Visual Fusion in a Noisy Recognition Task », in Proc. of 1st IEEE Workshop on Multimedia Signal Processing, Princeton, New Jersey, USA, 23-25 June 1997, pp. 37-44.

Reconnaissance audiovisuelle de voyelles bruitées : Estimation de la fiabilité du capteur audio et contrôle du processus de fusion

Pascal Teissier (1) (2), Jean-Luc Schwartz (1) et Anne Guérin-Dugué (2)

(1) Institut de la Communication Parlée CNRS UPRESA 5009 / INPG - U. Stendhal
ICP, INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex 1 / (teissier, schwartz)@icp.inpg.fr
(2) Laboratoire Images et Signaux
LIS, INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex 1/ guerin@tirf.inpg.fr

ABSTRACT

In order to fully exploit the auditory and visual complementarity, and realize an audiovisual speech recognition system robust in noise, we have to insert in the sensor fusion process a *contextual* information about the sensors *reliability*. Here, we show how to efficiently estimate the acoustic sensor reliability thanks to a non-linear preprocessing by "Curvilinear Component Analysis", and the way to insert this estimated context into the fusion process.

1. INTRODUCTION

Depuis plusieurs années, un certain nombre de systèmes de reconnaissance automatique de la parole intègrent une entrée visuelle de façon à améliorer l'identification de la parole dans le bruit acoustique [Sto96]. Le challenge consiste alors à obtenir la meilleure synergie : le taux de reconnaissance audiovisuelle doit être plus élevé que les scores auditif et visuel séparément comme cela est constaté pour des sujets humains. Quelle que soit l'architecture de fusion employée, il semble nécessaire, pour réaliser ce challenge trop rarement atteint, d'introduire une information (contexte) afin de guider le système vers la modalité la plus efficace [Sch98] [Mov96]. De nombreux travaux vont dans ce sens, cependant le contexte est très rarement estimé : il est souvent calculé à partir d'une information fournie sur le niveau de bruit.

Le présent travail porte sur l'estimation et l'introduction de contexte sur un modèle de fusion audiovisuelle pour une application à la reconnaissance de voyelles audiovisuelles du français à différents rapports signal sur bruit (RSB).

2. CONDITIONS EXPERIMENTALES

2.1 Données audiovisuelles

Le corpus comporte 100 répétitions de chacune des 10 voyelles orales du français [i, e, ɛ, y, ø, œ, u, o, ɔ, a] prononcées isolément par un seul locuteur. Le corpus a été enregistré avec un poste "visage parole" [Lal90] qui nous permet d'extraire trois paramètres caractéristiques du contour labial, l'étirement (A), la hauteur (B) et la surface (S) du contour interne des lèvres qui fournissent les trois entrées vidéo. Les signaux acoustiques bruités sont obtenus en additionnant un taux variable de bruit blanc gaussien avec 8 rapports signal sur bruit : sans bruit, 24 dB, 12 dB, 6 dB, 0 dB, -6dB, -12dB et –24 dB. Les entrées audio sont 20 coefficients spectraux en sortie d'une analyse spectrale utilisant une échelle perceptuelle en bark pour les fréquences (voir [Tei97] pour plus de détails).

2.2 Corpus d'apprentissage et de test

On considère ici le paradigme "d'extrapolation" pour lequel le corpus d'apprentissage contient uniquement des échantillons à fort RSB (sans bruit, 24 dB, 12 dB, 0 dB) et le système est ensuite testé avec des échantillons non appris sur tous les RSB. Nous avons utilisé dix partitions différentes de façon à améliorer la qualité des estimations des scores du système.

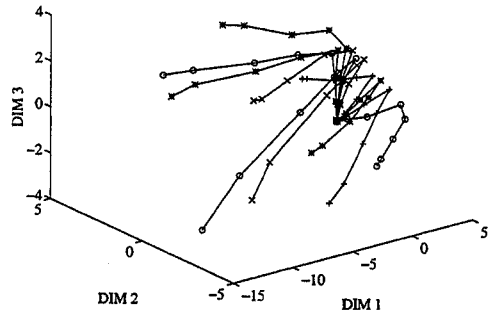
3 ARCHITECTURES DE FUSION AUDIOVISUELLE

Dans des travaux précédents, nous avons défini quatre architectures possible pour la fusion audiovisuelle [Sch98] et montré [Tei97] comment introduire dans chacune de ces architectures des principes de fusion de divers types, selon une taxinomie proposée par I.Bloch [Blo96]. Nous nous intéressons ici à une architecture classique, le modèle à Identification Séparée (IS) avec fusion dépendant du contexte.

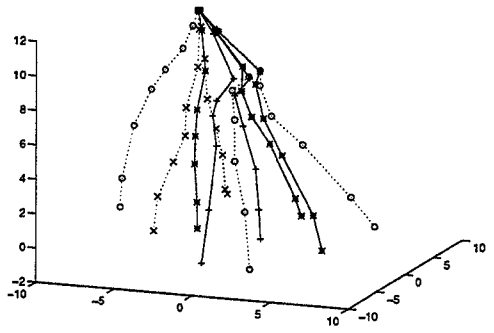
Dans ce modèle, les données acoustiques et visuelles sont identifiées séparément, puis on applique un processus de fusion de décisions pour estimer le score audiovisuel. Pour chaque classifieur monomodal, nous utilisons l'analyse discriminante quadratique (classifieur gaussien) : pour une partition donnée nous estimons la moyenne m_i et la matrice de covariance V_i pour chaque classe ω_i (avec 10 classes), puis nous calculons la probabilité *a-posteriori* $P(\omega_i/x)$ pour un vecteur d'entrée x donné. On appelle respectivement P_A et P_V les probabilités à la sortie du classifieur Audio et Visuel, et P_{AV} la probabilité à la sortie du processus de fusion. Lorsque le contexte n'est pas introduit P_{AV} est calculé grâce à un processus Bayesien classique. On introduit le contexte par le biais de facteurs de pondération exponentiels α et $(1-\alpha)$ qui renforcent sélectivement le poids des décisions auditive et visuelle dans le processus de fusion multiplicative :

$$P_{AV}(\omega_i/\vec{x}_{av}) = \frac{[P_A(\omega_i/\vec{x}_a)]^{\alpha_i} \cdot [P_V(\omega_i/\vec{x}_v)]^{1-\alpha_i}}{\sum_{j=1}^{10}[P_A(\omega_i/\vec{x}_a)]^{\alpha_j} \cdot [P_V(\omega_i/\vec{x}_v)]^{1-\alpha_j}} \quad (1)$$

La décision est fournie par *argmax* $P_{AV}(\omega_i/x)$.



(a)



(b)

Figure 1 – Représentation des trajectoires des centres de classe pour les 8 RSB à 3 dimensions après une projection par ACP (a) et ACC (b)

4. PRETRAITEMENT NON LINEAIRE DES DONNEES AUDITIVES

Une difficulté pour la classification audio et l'estimation de la fiabilité de ce classifieur est que dans l'espace auditif, les trajectoires produites par la variation des spectres vocaliques dans une catégorie donnée lorsque le bruit augmente sont très complexes. Une Analyse en Composantes Principales (ACP, prétraitement linéaire) réalisée sur tout le corpus confirme la complexité des trajectoires produites par la déformation du spectre de voyelle quand le bruit augmente (voir figure 1.a, chaque symbole représente le centre d'une classe à différents RSB). L'objectif du prétraitement non-linéaire est de simplifier les trajectoires des stimuli quand le bruit augmente (trajectoire "dépliée") de façon à faciliter la classification et l'estimation du contexte. Pour cela, nous utilisons une transformation réalisée par un réseau de neurones artificiel appelée "Analyse en Composantes Curvilinéaires" (ACC, [Dem97]) pour essayer de déplier ces trajectoires. La figure 1.b nous montre la projection en 3D du corpus audio en sortie de l'ACC apprise sur les niveaux de bruit d'apprentissage définis en section 2.2. Nous voyons clairement le dépliage des trajectoires et le rétrécissement de l'espace des voyelles (de bas en haut) dû au bruit. De plus, le processus d'ACC a été adapté de façon à faire émerger clairement la dimension de bruit *DIM 3* [Gue97]. Nous allons voir dans les prochaines sections l'intérêt potentiel du dépliage pour l'estimation et l'introduction d'un contexte.

5. EXPERIENCE 1 : INFLUENCE DU CONTEXTE

Dans cette première expérience, on compare, avec ou sans prétraitement non-linéaire des données auditives, les scores de reconnaissance selon l'introduction du contexte. On suppose que la valeur du RSB est connue, et on calcule l'information contextuelle α avec la formule suivante :

$$\alpha(x) = \frac{Power(Signal)}{Power(Signal)+Power(Noise)} = \frac{10^{\frac{SNR(x)}{10}}}{1+10^{\frac{SNR(x)}{10}}} \quad (2)$$

On peut aussi utiliser des formules α(SNR) plus complexes, avec processus d'optimisation [Tei97]. La figure 2 décrit les taux de reconnaissance pour des données auditives brutes (20D) et avec prétraitement non-linéaire (ACC-3D), selon quatre conditions : score visuel, score auditif, score audiovisuel sans contexte et score audiovisuel avec contexte (connu *a priori*).

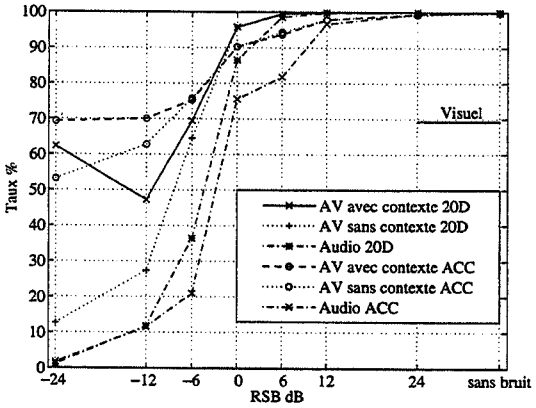


Figure 2 – Résultats de l'expérience 1

5.1 Cas sans prétraitement

Le score audiovisuel sans contexte chute considérablement en dessous de –6 dB. Le canal vidéo n'apporte que très peu d'amélioration sur la robustesse de ces modèles dans du bruit acoustique.

L'introduction d'une entrée contextuelle liée au RSB améliore sensiblement les taux de reconnaissance. Cependant le challenge n'est pas complètement obtenu. En effet, le poids de la composante P_A dans le processus de fusion (Eq. 1) reste trop fort à faible RSB, même pour les valeurs très faibles de α.

5.2 Cas avec prétraitement

Malgré un score audio légèrement inférieur, le score de reconnaissance sans contexte est sensiblement amélioré par rapport au cas sans prétraitement. Grâce au dépliage des

données par ACC, les trajectoires des voyelles dans le bruit sont assez rectilignes, et le classifieur audio fournit à faible RSB des probabilités *a posteriori* beaucoup moins contrastées, par rapport au cas sans prétraitement. Le poids des décisions audio dans le processus de fusion est alors naturellement atténué, d'où une fusion plus efficace.

Lorsque l'on introduit le contexte, on remarque que le challenge est atteint : le score audiovisuel reste supérieur ou égal au score visuel. La comparaison avec le cas sans prétraitement montre que le prétraitement non-linéaire permet une introduction efficace d'une entrée contextuelle liée au RSB : les probabilités a posteriori du classifieur audio, beaucoup plus ambigües à faible RSB, "pèsent" moins lourd dans la fusion.

Cette première expérience montre deux choses. Premièrement, la composante auditive bruitée joue un rôle trop important sur la décision finale, au détriment du comportement global de la reconnaissance audiovisuelle, et le prétraitement non-linéaire permet de diminuer sensiblement ce problème. Deuxièmement, une entrée contextuelle doit être systématiquement introduite de façon à guider le système vers la modalité la plus efficace.

6. ESTIMATION DU CONTEXTE

L'idée de l'introduction de paramètres de contrôle du processus de fusion reliés à la fiabilité des capteurs auditif et visuel (contexte) a fait son apparition récemment dans la communauté de la reconnaissance audiovisuelle de la parole avec deux voies possibles : contexte estimé en *sortie* des processus de classification de chaque capteur [Adj96][Mov96] ou directement à partir d'une estimation du RSB sur les données d'*entrée* [Mei96]. Nous avons exploré ces deux voies, pour l'estimation de la fiabilité du capteur *auditif*.

6.1 Estimation en sortie du classifieur

Un classifieur fournit une estimation de la conformité d'un stimulus à un ensemble d'apprentissage : ceci peut servir de point d'entrée à l'estimation du contexte, par diverses mesures (entropie, probabilités *a-priori*, ambiguïté). L'estimation du contexte selon la probabilité *a-priori* d'observer un stimulus étant donnée une classe gagnante (distance du stimulus à la moyenne de la classe, pondérée par la matrice de covariance) s'avère inutilisable dans notre cas. En effet, notre modélisation gaussienne produit des effets de queue de distribution symétriques, qui peuvent conduire à des estimations de contexte similaires à très fort et très faible bruit. L'utilisation de "l'ambiguïté" consiste à calculer une information sur le *contraste* des probabilités P_A en sortie du capteur audio : si la décision du capteur audio est ambiguë, on met plus de poids sur le capteur visuel. Pour calculer l'ambiguïté du capteur audio, nous avons utilisé la formule suivante :

$$\alpha(x) = \frac{P_A(\omega_i / x) - P_A(\omega_j / x)}{P_A(\omega_i / x)} \qquad (3)$$

où ω_i est la classe gagnante et ω_j est la seconde classe gagnante du classifeur audio. Nous présentons sur la figure 3 l'estimation du contexte par ambiguïté et les scores de reconnaissance audiovisuelle après introduction de cette estimation selon 2 cas : sans prétraitement (dimension audio n_a=20) et avec prétraitement non linéaire (par ACC, n_a=3). On remarque que l'estimation du contexte sans prétraitement est catastrophique. Ceci est dû à la complexité des trajectoires des voyelles à travers le bruit : à fort niveau de bruit, l'une des classes audio est souvent gagnante avec une très faible ambiguïté. Le score audiovisuel après introduction du contexte estimé est alors aussi très mauvais. Dans le cas du prétraitement non linéaire, ce problème est largement diminué : l'estimation du contexte est beaucoup mieux reliée au RSB, grâce à une meilleure modélisation des données due au dépliage des trajectoires. La dynamique de cette estimation est néanmoins trop faible, elle ne permet donc pas d'obtenir notre challenge : le score audiovisuel reste inférieur au score visuel à fort bruit et peu différent du score sans contexte (voir Fig. 2). Ceci est probablement dû au fait que, quel que soit le prétraitement, l'ambigüité de chaque capteur est déjà prise en compte dans le processus de fusion multiplicative (plus le classifieur est ambigu, plus son poids dans l'Eq.1 est faible).

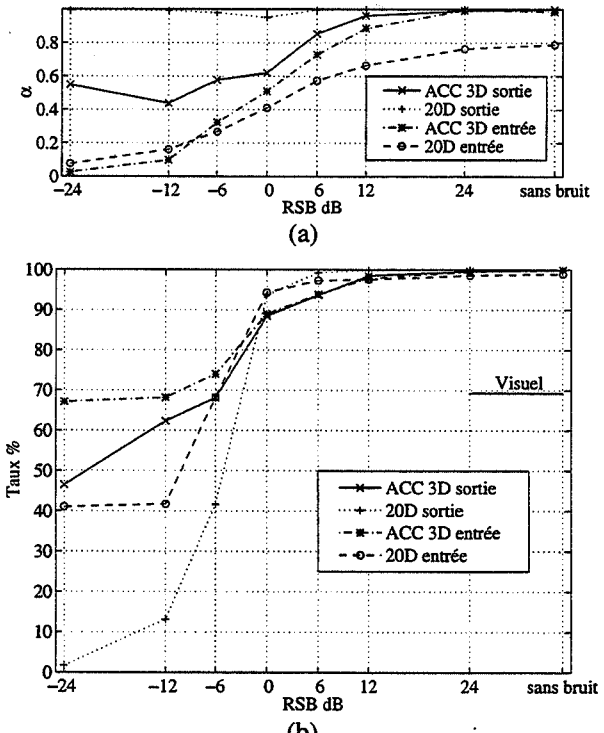


Figure 3 – Résultats de l'expérience 2 (a) Evolution de l'estimation moyenne pour les dix classes vocaliques de α en sortie ou en entrée du classifieur et pour deux types de prétraitement (ACC 3D et sans prétraitement 20D); (b) évolution des scores de reconnaissance audiovisuelle avec fusion dépendante du contexte pour ces deux types d'estimation et de prétraitement.

6.2 Estimation en entrée du classifieur

La position d'un échantillon sur la trajectoire des voyelles en fonction du bruit nous donne une information sur le RSB (donc sur la fiabilité du capteur audio).

A partir de l'entrée contextuelle théorique α (Eq. 2), pour chaque voyelle, nous calculons une estimation de α, $\hat{\alpha}_i$, obtenue par une régression polynomiale des données auditives définie sur les 4 niveaux de bruit d'apprentissage. Une fonction de régression est calculée pour chaque classe afin de prendre en compte les différences des trajectoires des voyelles dans le bruit. On utilise pour les données 20 D non prétraitées une régression linéaire, et pour l'ACC une régression quadratique sur la 3^{ième} dimension, très corrélée au bruit (Fig. 1b). C'est alors la valeur estimée qui est introduite dans le processus de fusion (Eq. 1).

L'évolution du facteur de pondération estimé $\hat{\alpha}$ est décrite sur la figure 3a. Sans prétraitement, on peut clairement remarquer que l'estimation du contexte n'est pas précise à cause de la complexité des trajectoires. Par contre, lorsque les trajectoires sont dépliées à l'aide du prétraitement non linéaire, l'estimation du contexte est tout-à-fait correcte. La figure 3b nous indique les scores de reconnaissance audiovisuelle après introduction du contexte estimé. Sans prétraitement, le score audiovisuel n'est pas satisfaisant à cause d'une mauvaise estimation du contexte : le poids du capteur vidéo est trop important à faible RSB et trop faible à fort RSB. Avec un prétraitement non linéaire, la bonne estimation du contexte permet une reconnaissance audiovisuelle efficace : le score audiovisuel avec contexte estimé est similaire au score audiovisuel avec contexte connu *a priori* (voir figure 2). L'objectif est atteint : le score de reconnaissance audiovisuelle converge vers le score de reconnaissance visuelle quand le RSB diminue et il reste supérieur ou égal aux scores visuel et auditif dans toute la dynamique de RSB.

7. Discussion et conclusion

Après avoir montré l'intérêt de l'introduction d'une information contextuelle, nous avons présenté deux façons d'estimer le contexte afin de prendre en compte la fiabilité du capteur acoustique pour une application à la reconnaissance audiovisuelle. La première technique est basée sur l'information d'ambiguïté de la décision du capteur acoustique, elle est très facile à implémenter mais la difficulté est d'adapter l'évolution de ce paramètre pour une pondération efficace de toutes les décisions à tous les niveaux de bruit. La deuxième technique exploite la structure intrinsèque des données. Pour cela, on utilise les trajectoires des échantillons dans une configuration bruitée. Pour ces deux techniques, le prétraitement par ACC est bien adapté puisqu'il permet une représentation quasi optimale en révélant la structure intrinsèque des données. Cette représentation facilite donc la modélisation des données et permet une estimation du contexte, et une reconnaissance audiovisuelle efficace par rapport à un prétraitement linéaire. En résumé, prétraitement non linéaire, estimation de la fiabilité du capteur auditif et fusion dépendante du contexte nous permettent de réaliser un système de reconnaissance audiovisuelle efficace à tous les niveaux de bruit.

8. Remerciements

Ce travail est supporté par la Fédération de laboratoires ELESA du CNRS-INPG.

9. References

[Adj96] Adjoudani, A. & Benoît, C. *"On the integration of auditory and visual parameters in an HMM-based ASR"*, In [Sto96] pp. 461-472, 1996.

[Blo96] Bloch, I. *"Information Combination Operators for Data Fusion : A Comparative review with Classification"*, IEEE Trans on SMC, A, vol 26, 1, pp. 52-67, 1996.

[Dem97] Demartines, P. & Hérault, J. *"Curvilinear Component Analysis : A Self-Organizing Neural Network for Non Linear Mapping of Data Sets"*, IEEE Trans on Neural Networks, Vol 8, 1, pp. 148-154, January 1997.

[Gue97] Guérin-Dugué, A. et al. *"Non linear representation for audio-visual fusion in a noisy-vowel recognition task"*, NEURAP'97, Marseille, pp. 31-40, March 1997.

[Lal90] Lallouache, M.T. *"Un poste visage parole. Acquisition et traitement de contours labiaux"*, XVIII Journées d'Etudes sur la Parole, Montréal, pp. 282-286, 1990.

[Mei96] Meier, U. et al. *"Adaptive bimodal sensor fusion for automatic speechreading"*, ICAPSS'96, 1996.

[Mov96] Movellan, J.R. & Mineiro, P. *"Modularity and catastrophic fusion : a bayesian approach with application to audiovisual specch recognition"*, Technical Report CogSci. USCD-97.01, Department of Cognitive Science, USCD, San Diego, CA, 92093-0515, 1996.

[Sch98] Schwartz, J.L. et al. *"Ten years after Summerfield... a taxonomy of models for audiovisual fusion in speech perception"*, In R. Campbell, B. Dodd & D. Burnham (eds.) Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing. Erlbaum/Psychology Press, pp. 85-108, 1998.

[Sto96] Stork, D.G.& Hennecke, M.E. *(Eds.) Speechreading by Man and Machine: Models, Systems and Applications*. NATO ASI Series, Springer, 1996.

[Tei97] Teissier, P et al. "Comparing models for audiovisual fusion in a noisy-vowel recognition task", Submitted to IEEE Trans. Speech and Audio Processing, 1997.

Phonétique II

L'étude des voyelles [*a*] et [a] en français et en persan

Assadi, Sh.

ILPGA-19, rue des Bernardins. 75005 Paris UA 1027

I. Introduction

Pendant l'acte de parole, le mouvement des différents articulateurs pour la production des segments phonétiques successifs interactent les uns sur les autres et la configuration du conduit vocal est influencée en même temps par plus d'un segment. L'effet acoustique de la coarticulation peut être observé par l'analyse spectrographique. La réduction de contraste entre la consonne et la voyelle subséquente semble être un phénomène universel, mais qui peut être différent selon les langues. Les changements phonétiques ont souvent commencé par les phénomènes combinatoires synchroniques. Mais la résistance à la coarticulation, n'est pas nécessairement un phénomène universel et varie selon le contexte et la langue [Bla, 1976]. Pour la production d'une séquence de consonne - voyelle (cv), le conduit vocal prend la forme requise pour la consonne, il est aussi influencé par la position appropriée pour la voyelle. La transition cv dépend de la trajectoire entre la consonne et la voyelle. Les valeurs des formants au début, au centre et à la fin de la voyelle montrent l'influence du lieu d'articulation des consonnes environnantes. Une constriction au milieu de la cavité buccale, est une position optimale pour un F2 élevé, et un F2 très bas nécessite une constriction au-dessus du pharynx [Fan, 1969]. Avant les voyelles antérieures /a, i, e/, /k/ et /g/ deviennent palatales et avant les voyelles postérieures /â, u, o/, ils sont vélaires.

Dans cette étude nous avons choisi la voyelle postérieure [*a*] et la voyelle antérieure [a]. Nous étudions l'effet de la coarticulation, la variabilité du timbre selon le contexte et l'effet d'hypo et d'hyperarticulation. Nous vérifions :

- Si les formants F2 et la distance F2-F1 permettent de distinguer les voyelles antérieure et postérieure dans les deux langues et si l'affaiblissement de la distinction entre les deux A en français d'aujourd'hui permet de confirmer des réalisations différentes .
- Si la variabilité du F2 selon le contexte, est plus importante dans une des deux langues.
- Si le trait antérieur/postérieur devient plus marqué en hyper articulation dans les deux langues.

Choix des voyelles /a/ et /*a*/.

1-Les statistiques sur la fréquence des voyelles montrent que /a/ est une des voyelles la plus fréquente dans les deux langues, pour le français, le % est de 17,47 ([Wio, 1972] concernant le français parlé), et pour le persan de 22,65 ([Gin, 1965], dans un mélange du style, familier et écrit.

2- Le /a/ est une des voyelles la plus variable du point de vue acoustique en français ([Vai, 1987], étude de seize combinaisons de consonnes et de voyelles du type CVCVCV, prononcées à un seul débit), et du point de vue réalisation allophonique individuelle, en persan [Pis, 1985].

3- L'étude diachronique des deux langues indique qu'autrefois la distinction des deux voyelles aurait été liée à la durée.

4- La distinction entre ces deux voyelles est rentable en persan (la preuve en est l'existence de beaucoup de paires minimales à l'initiale et en médiane). Mais elle tend à disparaître dans le français d'aujourd'hui [Del, 1966]. La distinction dépend de l'âge, de l'appartenance géographique [Mar, 1941] et du milieu social [Met,1979]. Les deux A s'acheminent vers un seul en français. Ceci est important pour les apprenants persans, qui ont deux A distincts dans leur langue, et pour parler le français cherchent une norme, mais cette norme existe-elle?

II. Procédure expérimentale

En débit normal, les locuteurs ont prononcé le corpus selon leur rythme habituel, en hypo et puis hyperarticulation, nous avons demandé aux locuteurs de prononcer les logatomes une fois d'une façon très relâchée et une fois le plus clairement possible (imitant une mère parlant à son enfant).

II.1- Le corpus: Le corpus est constitué de séquences de CVCV'CV accentuées sur la dernière syllabe. C= [b, d, g] et V= [*a*, a]. Les consonnes occlusives sourdes étant aspirées en persan, nous avons préféré les sonores aux sourdes. Les logatomes ont été introduits dans des phrases cadres du type « dis.... deux fois ». « *bogou.....do bâr* ». Nous avons choisi la phrase cadre à l'impératif, car dans ce cas, sa structure dans les deux langues étudiées, est identique, tandis que l'ordre d'une phrase déclarative en persan est : sujet + objet + verbe.

II.2- Les locuteurs : Les logatomes ont été prononcés par six sujets, quatre français (parisiens) et deux persans

(chirazis[1]), tous du milieu universitaire, et âgés de 40 à 50 ans. Une enquête auprès de 100 parisiens, nous a permis de sélectionner des locuteurs français faisant encore la différence à différents degrés entre les deux voyelles dans les paires minimales. Afin que les locuteurs prononcent correctement les logatomes avec la voyelle désirée, nous leur avons rappelé la distinction qu'ils font entre 'patte' et 'pâte' (pour les français) et 'bad' et 'bâd'(pour le persan).

I.3- les mesures: les formants F1 et F2 de la deuxième voyelle (CV du milieu, non accentué), au début et au milieu de la voyelle, ainsi que la distance entre F2-F1 ont été mesurés. Le degré de coarticulation entre la consonne et la voyelle est consideré comme la différence du F2 du début et du milieu de la voyelle (une petite différence indique beaucoup de coarticulation)

II.4- L'appareil de mesures: l'analyse acoustique est réalisée à l'aide d'un logiciel de traitement du signal UNICE.

III. Analyse des données

III.1- L'effet de la coarticulation

III. 1. 1 L'étude inter langue

La coarticulation semble être différente surtout dans le contexte /g/, due à l'effet de l'anticipation plus grande de la voyelle en persan : la différence de F2 entre le début et le milieu de la voyelle est moindre, (figure1). La voyelle et la consonne seraient donc plus coarticulées en persan (une petite distance indique moins de mouvements articulatoires, et par conséquence plus de coarticulation). En français, à l'exception d'un seul locuteur, les autres utilisent la seule variante palatale de /g/ avant /*a*/ et /a/, qui confirme le rapprochement des deux A. Les persans utilisent la variante palatale devant /a/, et la variante vélaire devant /*a*/. Les formants F2, ainsi que la distance entre F2-F1 de la voyelle antérieure (au début et au milieu de la voyelle) sont plus élevés en persan qu'en français. Une grande distance de F2-F1 est le corrélat d'un avancement de la langue et une distance petite, du recul de la langue, donc /a/ persan est plus antérieur. Cette différence acoustique permet une plus grande contrastivité phonologique entre les deux voyelles en persan. Dans le contexte labial, il y a peu de mouvements de formants dans les deux langues.

III. 1. 2- L'étude intra langue

La distance F2-F1 (au début, au centre de la voyelle) est influencée par le lieu d'articulation des consonnes environnantes (b<d<g). Cependant on constate que F2 de [g*a*] chez les deux locuteurs persans est moins élevés que F2 de [d*a*] (figure 1). Cela semble indiquer qu'en persan, le trait postérieur se propage de la voyelle sur la consonne, alors qu'en français, cette propagation n'a pas lieu. La différence inter locuteurs est importante en français. Un des locuteurs a produit les voyelles postérieure et antérieure semblables à celles du persan.

III. 2- La variabilité

III.2.1- L'étude inter langue

La variabilité du F2, selon le contexte, au centre des deux voyelles est plus importante, en persan qu'en français. La variation de F1 selon le contexte est minime, chez tous les locuteurs des deux langues.

III. 2. 2- L'étude intra langue

Dans le contexte de la voyelle antérieure ([ba, da, ga]), La variabilité est plus importante que dans celui de la voyelle postérieure ([b*a*, d*a*, g*a*]), chez tous les locuteurs des deux langues.

III. 3- L'effet d'hypo et d'hyperarticulation

III. 3. 1- Etude inter langue

En cas d'hyperarticulation, en général, F1 augmente chez tous les locuteurs des deux langues (renforcement du trait + ouvert). La voyelle antérieure est antériorisée à différents degrés dans les deux langues, (figure 2a et 2b), mais la voyelle postérieure est davantage postériorisée en persan. La distance entre F2-F1 de [a] est plus grande en persan dans tous les contextes.

III. 3.2- Etude intra langue

En demandant aux locuteurs de prononcer le plus clairement possible, ils ont tendance à utiliser des efforts et des stratégies différentes: il semble qu'ils insistent, soit, sur la voyelle, soit, sur la consonne, car il est difficile de contrôler le degré d'hyper aticulation. En français, en hyper articulant, un seul locuteur, qui faisait une distinction claire entre les deux voyelles, a postériorisé davantage la voyelle postérieure (abaissement du F2). Par contre, en persan, nous constatons cette tendance chez les deux locuteurs.

[1] Chiraz est une ville au sud de l'Iran.

IV-Conclusion

Les sons que nous prononçons sont le résultat d'un compromis entre la tendance assimilatrice et la nécessité de se faire comprendre. La coarticulation est un phénomène universel, qui peut varier selon la langue, le contexte et le locuteur. Nos résultats semblent refléter des différences de coarticulation surtout en contexte /g/, qui peuvent être dues à l'effet d'anticipation plus grand de la voyelle sur la consonne en persan. En français le degré de la coarticulation semble varier selon le sujet. Un des locuteurs a réalisé la voyelle postérieure et antérieure semblables à celles du persan. Les autres locuteurs français, qui lors du questionnaire, ont prouvé qu'ils distinguaient de façon régulière entre /a/ et /a/, ont prouvé une certaine instabilité dans leurs productions et ils n'ont pas la stratégie nécessaire pour maintenir la distinction dans les non mots. Par exemple, ils utilisent la seule variante palatale de /g/ avant /a/ et /a/, alors que les persans utilisent la variante palatale devant /a/ et la variante vélaire devant /a/, le contraste antérieur / postérieur est moins évident dans les non mots en français. La distance entre F2-F1 de [a] en persan est plus grande que celle du français, le [a] persan est donc plus antérieur. La variabilité du F2 selon le contexte, est plus importante en persan.

En général, en hyperarticulation, le degré d'aperture augmente dans les deux langues, mais le trait antérieur/postérieur devient plus marqué en persan. La différence entre F2-F1 de la voyelle antérieure est plus grande en persan qu'en français.

Les expériences du laboratoire comme celle-ci conduisent les locuteurs à parler avec une certaine stratégie, mais nous donnent des renseignements intéressants sur les différents aspects d'une langue.

A l'exception des variabilités inter locuteur dont les causes sont les dimensions et les proportions de l'appareil vocal, les habitudes articulatoires de chacun peuvent varier les fréquences des formants. Mais la variabilité du signal de la parole, qu'elle soit la variabilité inter ou intra locuteur, ou, inter langue confirme la propriété dynamique du mécanisme articulatoire.

V-Bibliographie

[Bla. 1976]

Bladon, R. A.W.& AL BAMERNI, A. (1976) « Coarticulation resistance in English /l/. » Journal of phonetics,. 4, pp. 137-150.

[Del. 1966]

Delattre, P. (1966) Studies in French and Comparative Phonetics. Mouton & Co. The Hague .

[Fan.1969]

Fant G. (1969) Stop in CV syllables. Quarterly Progress and status Report, KTH. Speech Transmission Laboratory.

[Fat.1960]

Fant, G.(1960) Acoustic theory of speech production. Mouton. The Hague.

[Kru.1987]

Krull, (1987) « Second formants locus patterns as a measure of consonant - vowel coarticulation », Perilus V, pp.43-61. Institute of Linguistics, University of Stockholm.

[Mar. 1941]

Martinet, A. (1941) La prononciation du français contemporain. Genève.

Met. [1979]

Mettas, O. (1979) La prononciation Parisienne. Aspect phonétique d'un sociolecte Parisien du Faubourg Saint Germain à la Mouette. France.

[Moo.1994]

Moon, S-J. and Lindblom, B. (1994) « Interaction between Duration, Context and Speaking Style in English Stressed Vowels ». J .A.S.A. 96 *(1) pp. 40-55.*

[Pis.1979]

Pisowicz, A. (1979) Origine of the new and middle Persian. Phonological system. Krakow.

[Vai. 1987]

Vaissière, J. (1987) The uses of allophonic variation of /a/ in automatic continuous speech recognition of French. Speech communication group working paper. M. I. T. Reseaerch Laboratory of Electronics.15- 25.

[Wio.1972]

Wioland, (1972), Estimation de la fréquence en français parlé. Travaux de l'institut de phonétique de Strasbourg. 4. pp.177-204.

Figure 1: La moyenne des formants F2 de tous les locuteurs de chaque langue, dans les différents contextes (les mesures sont prises au début et au centre de la voyelle). (pour faciliter, dans cette étude, *a* est présenté comme â).

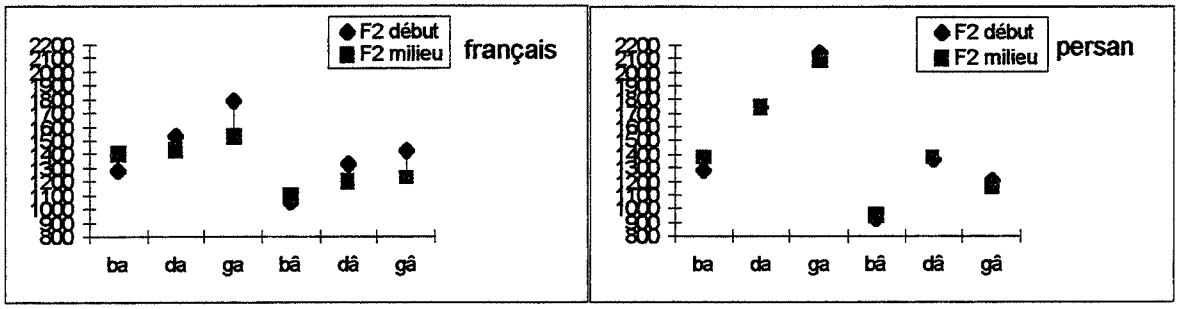


Figure 2a: Les formants F1 et F2 en hypo et hyperarticulation en français (pour chaque locuteur, les mesures sont prises au centre de la voyelle).

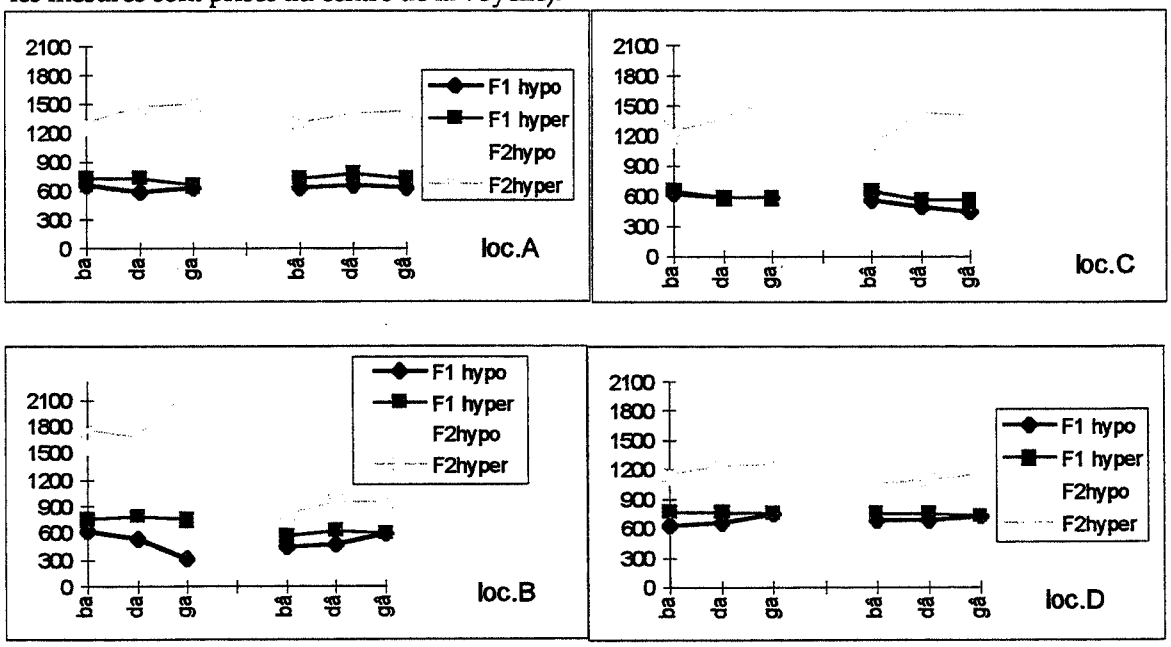
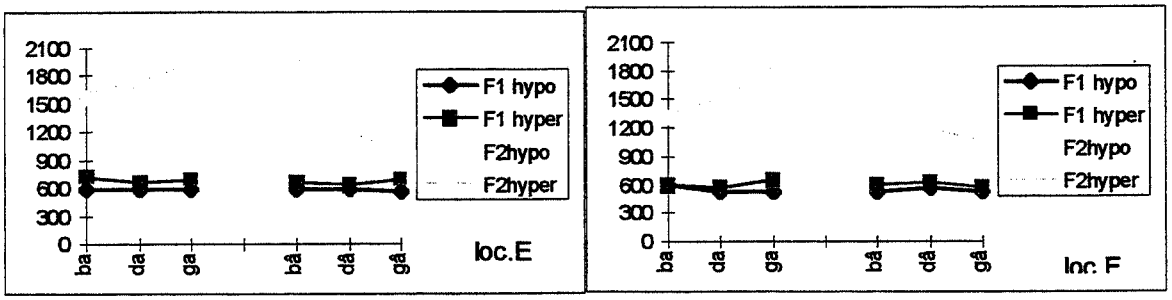


Figure 2b: Les formants F1 et F2 en hypo et hyperarticulation en persan (pour chaque locuteur, les mesures sont prises au centre de la voyelle).



VARIABILITE INTRA-LOCUTEUR DANS LA PRODUCTION DU PHONEME /R/ EN FRANCAIS

Michel CHAFCOULOFF et Denis AUTESSERRE

Laboratoire 'Parole et Langage', ESA 6057 CNRS
29, Av. R. Schuman, 13621, Aix-en-Provence, France

Abstract

To investigate the intra-speaker variability in the production of the [R]-sound in Southern French, several factors of variation (position, vocalic context, idiolectal factors) have been considered. Results of an acoustical analysis show that several allophones of /R/ were produced, i.e. a single or double flapped allophone in the initial position, an approximant and/or a voiced constrictive in the medial position, a devoiced constrictive in the final position. Particular attention was paid to the direction of formant transitions as a function of contextual environment. As a conclusion, it results that the combination of these variation factors contribute to establish [R] as a highly coarticulated, polymorphic sound in this speaker's dialect.

1. Introduction

La transcription phonétique des principaux allophones du phonème unique /R/ en français en utilisant l'alphabet phonétique international (API) ne devrait recourir qu'à des caractères diacritiques additionnels venant modifier la lettre choisie pour représenter le phonème. Or, dans ce cas, (et il n'est pas le seul) le principe phonémique sous-jacent à ce code de transcription, n'est pas appliqué et il est habituel de recourir à deux caractères différents, pour noter les principales variantes de ce phonème. Dans la représentation sous forme de tableau de l'API, ceci est dû au fait que ces allophones, appartiennent à des classes différentes de mode d'articulation : l'un étant battu [R], l'autre constrictif [ʁ]. En revanche, l'unicité est retrouvée dans ce même tableau par l'attribution d'un lieu d'articulation uvulaire unique. Mais lorsque l'on connaît la sensibilité à la coarticulation de cette consonne, on peut se poser la question de savoir s'il est judicieux de regrouper toutes ces formes dans une macro-classe de lieu d'articulation. En fait, la réalisation phonétique [R] est-elle inclassable ou plutôt a-t-elle été mal classée ? Pour apporter des éléments de réponse à cette question, une analyse acoustique fine des différentes variantes de /R/ a été effectuée chez un même locuteur.

2. Protocole expérimental

2.1 Constitution et enregistrement du corpus

Comme l'objet de cette étude porte plus particulièrement sur la variabilité de /R/, plusieurs facteurs de variation (positionnelle, contextuelle, individuelle) ont été pris en compte dans l'élaboration du corpus. Celui-ci est composé de mots lexicaux où la consonne apparaît en position initiale, médiane et finale en contexte vocalique [i],[a],[u]. En ce qui concerne les consonnes [R] en position initiale et finale, les mots ont été placés dans des phrases courtes alors que pour la position intervocalique (en contexte symétrique), on a utilisé une phrase porteuse du type 'Il faut répéter le mot....deux fois'. Chaque série de mots a été enregistrée à trois reprises dans une chambre anéchoïque par un locuteur représentatif d'un idiolecte méridional peu marqué.

2.2 Analyse acoustique

L'analyse acoustique a été fondée sur l'utilisation du logiciel 'Phonedit'. Le signal de parole, après numérisation, est affiché sur un écran d'ordinateur et est représenté sous une double forme oscillographique et spectrographique. La mesure des fréquences des cibles acoustiques et des transitions de formants, ainsi que celle de l'intensité relative entre les pics d'énergie a été effectuée par FFT à des fins de représentation rigoureuse du suivi des trajectoires formantiques.

2.3 Présentation des données acoustiques

Il convenait de rendre compte d'une manière globale du comportement des indices acoustiques en fonction des différentes sources de variation. Pour les voyelles environnantes, les valeurs des formants ont été choisies à partir des changements de direction des transitions influencées par la consonne [R]. Ces mêmes critères ont été réutilisés dans la représentation des pentes et de la direction des transitions formantiques. En ce qui concerne la variante battue à l'initiale absolue, l'appui

vocoïdal a été restitué dans sa durée intégrale. Selon les variantes considérées, on a pris en compte d'une part les formants pour les approximantes ou les battues, d'autre part les bruits pour les constrictives. Les principaux indices extraits de l'analyse acoustique sont représentés de façon schématique sur la figure 1. Les transitions apparaissent en traits pleins, alors que les trajectoires formantiques traversant la consonne sont en pointillés.

3. Variabilité intra-individuelle

La consonne [R] est toujours placée en syllabe accentuée dans des positions où elle se trouve directement en contact avec une voyelle qui précède et/ou qui suit. Dans ces conditions, la variante battue se trouve surtout à l'initiale absolue, comme par exemple dans la phrase 'Rares sont les vrais amateurs'

3.1 Positions initiale et intervocalique

Ainsi positions intervocaliques et initiales de /R/ sont rapprochées puisque dans les deux cas la consonne apparaît précédée et suivie d'éléments vocaliques, plus précisément un subsegment vocoïdal pour la variante battue.

Contextes vocaliques [u] et [a]

L'amplitude des variations formantiques est tout aussi faible pour les variantes battues et approximantes. La vocoïde est alors proche par son timbre de la voyelle, et la consonne ou le battement se manifestent de la même manière par une aggravation et une chute d'intensité de tous les formants. Par application de l'équation de locus de F2, [Chaf. et Aut.97a], nous avons démontré que la résistance coarticulatoire de la consonne aux effets de la voyelle est faible en contexte [a] et [u].

Contexte vocalique [i]

La situation est totalement différente. L'attaque vocoïdale de la variante battue en position initiale absolue se réalise dans le grave avec un timbre proche de celui de la voyelle [o]. Pour l'élément contoïdal, seul le premier formant présente une aggravation de la fréquence. Une importante remontée du F2 s'accompagne de transitions positives tant pour la vocoïde que pour la voyelle subséquente. Le troisième formant présente une situation intermédiaire de fléchissement dans la première partie suivi d'une élévation vers la voyelle subséquente. Ce mouvement général du F3 est précédé de transitions négatives sur l'élément vocoïdal et positives pour la voyelle, les transitions devenant positives sur la voyelle. Cette fois-ci, le rapprochement entre la position initiale et l'intervocalique est moins fondé, surtout pour ce qui est du deuxième formant. La voyelle placée avant la consonne a une transition inverse de celle qui était réalisée avec la vocoïde. Pendant la réalisation de la consonne intervocalique, l'évolution de la structure formantique prend davantage l'allure d'une vallée plus creusée que dans le cas des voyelles [u] et [a]. Les résultats de l'étude citée plus haut, avaient déjà montré une plus grande résistance à la coarticulation en contexte [i]. Ce fait trouve son illustration, plus particulièrement en position initiale où l'on on constate que l'influence de cette voyelle se manifeste par des trajectoires abruptes importantes (de l'ordre de 500Hz) qui concernent le deuxième et le troisième formant de la consonne.

Comparaison entre les contextes vocaliques

Les trois réalisations de [R] dans un même contexte vocalique montrent aussi un comportement différent en fonction du type de voyelles adjacentes. Pour les trois répétitions de la consonne en contexte vocalique [u] et [a], les valeurs de formants restent très proches. Dans ces cas, la dépendance vocalique est très forte et la consonne apparaît comme un glissement avec une légère aggravation d'une voyelle à l'autre [Öhm.66].
A la variation importante des fréquences de formants sur la consonne, s'ajoute en contexte vocalique [i], l'apparition de bruits (bien visibles dans la première occurrence en position intervocalique). Ces derniers sont présents dans une région située un peu plus haut que la cible formantique de la variante approximante. Entre 1700 et 1800 Hz pour le deuxième formant et entre 2500 et 2700 Hz pour le troisième formant. Il y a donc dans le contexte intervocalique [iRi] possibilité d'une réalisation soit approximante, soit constrictive.

3.2 Position finale

[R] consonne pharyngale ?

Par rapport aux deux positions déjà étudiées, on remarque en finale une inversion de pente des transitions. Cette pente ascendante de F1 a été interprétée par Delattre [Del69] comme un indice de pharyngalisation. De notre point de vue, cette opinion n'est justifiée ni sur le plan acoustique, ni sur le plan articulatoire pour les voyelles [i] et [u]. L'observation de films cinéradiographiques [Roc71] et [Bot86] montre que le passage pharyngé reste largement ouvert dans ces deux derniers contextes. D'autre part, le recul de la langue dans la région uvulaire occasionne une remontée de la transition de F1 et une baisse concomitante de F2 [Sant.72]. Les valeurs des fréquences terminales de la transition vocalique de F1, se trouvent légèrement au dessous de 500 Hz pour [i] et [u], et aux alentours de 800 Hz pour [a]. Ce qui conduit à interpréter [R] comme une consonne vélo-uvulaire en contexte [i] et [u] et pharyngale en contexte [a].

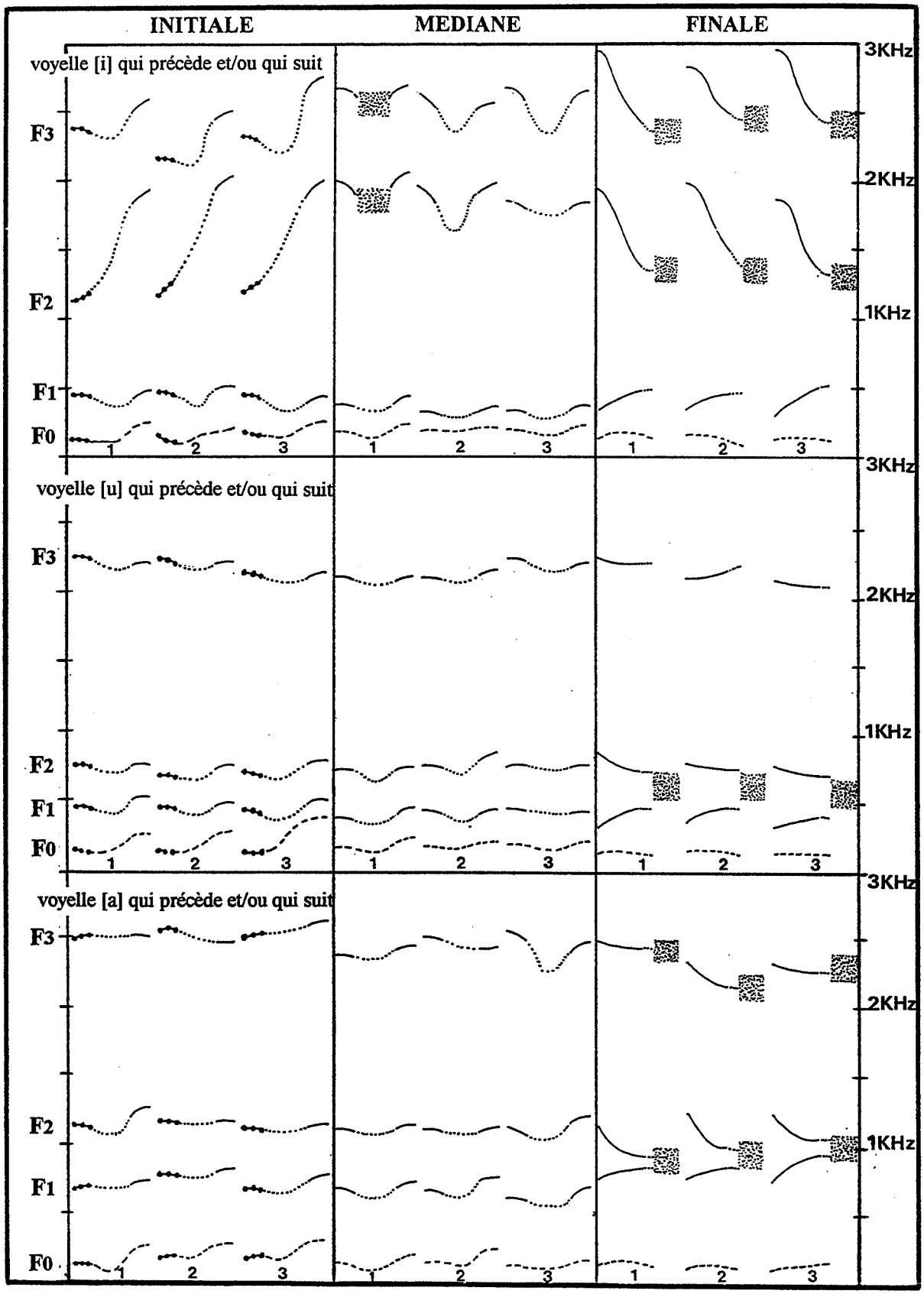


Figure 1 : Principaux indices acoustiques des variantes de [R] en position initiale, médiane (intervocalique), finale et en contexte vocalique [i], [u], [a].— Transitions vocaliques ; .... Trajectoires formantiques sur la consonne ; - - - Variations de la fréquence fondamentale (F0), ▦ Zones de bruit , ●●● Attaque vocoïdale

Dévoisement, bruit et transitions

Le dévoisement est la règle générale en position finale avec pour caractéristique acoustique majeure la présence de zones de bruit d'intensité plus marquée. La plus grave (500-700 Hz) apparaît pour la voyelle [u]. En ce qui concerne [i] et [a], deux zones de bruit ressortent : la première de fréquences moyennes (900-1200 Hz pour [a], la deuxième de fréquences plus élevées (2300-2600 Hz) pour les deux voyelles [i].
La transition de F1 est positive et celle de F2 négative dans les différents contextes vocaliques étudiés. On constate même un rapprochement, voire une fusion des transitions de F1 et de F2 qui convergent vers la zone de bruit grave pour [u] ou moyenne pour [a].La direction de ces transitions est identique à celle prédite dans le modèle à 8 régions distinctives de Mrayati et al. 1988), [Mra88]. Ceci implique que la consonne [R] a été articulée dans la région C (région pharyngale) du conduit vocal délimitée par ces auteurs.
Un même indice acoustique, la remontée de F1 correspond bien à une réalisation pharyngale de [R] en contexte [a] où elle atteint une valeur de 800 Hz .En revanche, elle demeure uvulaire en contexte [i] et [u] où les valeurs formantiques restent aux alentours de 500 Hz. Pour la variante constrictive partiellement dévoisée, on remarque en contexte vocalique [i] une évolution de la transition de F2, mais cette fois-ci sur la voyelle qui précède inverse de celle que l'on avait en position initiale, mais sur la partie contoïdale de la variante (dite battue),

4. Conclusion

Cette analyse acoustique nous a permis de mettre en évidence quatre variantes principales du phonème /R/ en français. La variante dite battue (le plus souvent à battement unique) se rencontre plutôt à l'initiale de l'énoncé ou dans un groupe consonantique. Elle constitue un héritage des contraintes phonotactiques d'occurrence de la variante battue alvéolaire qui l'a précédée dans la langue. C'est surtout la présence d'une vocoïde subsegmentale sur la trajectoire uvulaire, qui permet de définir cette variante. En position médiane intervocalique, le battement isolé a disparu. Il est remplacé par une constriction plus ou moins marquée, qui donne lieu à la réalisation de variantes approximantes ou constrictives (faiblement bruitées). La variante vraiment perçue comme fricative, apparaît surtout en position finale d'énoncé où l'interruption du voisement favorise la prééminence de la source de bruit. Il semble difficile de rechercher des invariants tant dans le domaine du mode que du lieu d'articulation. Si l'on doit constater l'existence d'un élément commun, il faudrait le trouver dans le comportement de l'organe articulateur, et plus particulièrement dans la rétraction du dos de la langue qui se dirige vers la cible uvulaire ou pharyngale en fonction du contexte vocalique. Dans un environnement de voyelles hautes comme [i] et [u], la rétraction linguale entraîne la langue vers la luette, mais laisse le pharynx libre, en particulier dans le contexte symétrique antérieur. L'abaissement de la langue pour la voyelle [a] entraîne par contrecoup sa rétraction, d'où la réalisation d'une constriction pharyngale associée à une anticipation de l'abaissement pour la voyelle. Ainsi, la consonne [R] apparaît dans cet idiolecte comme résultant d'un geste articulatoire de rétraction linguale, et on pourrait en faire sur le plan phonologique le trait distinctif caractéristique de cette consonne. La sensibilité particulière à la coarticulation, ne nous a pas permis de dégager des invariants acoustiques absolus [Öhm97].

BIBLIOGRAPHIE

[Aut. et Chaf.97a] Autesserre, D. et Chafcouloff, M.(1997a) 'Le sub-segment vocalique associé à la réalisation d'un allophone de /R/ en français', *Actes des Journées d'Etudes Linguistiques, La voyelle dans tous ses états*, Nantes (5-6 décembre), 112-117

[Bot86] Bothorel, A., Simon, P. Wioland, F., Zerling, J.P. (1986) ' *Cinéradiographie des voyelles et consonnes du français'*, Travaux de l'Institut de Phonétique de Strasbourg , 287 p.

[Chaf. et Aut.97b] Chafcouloff, M. et Autesserre, D. (1997b) 'Examen fibroscopique de /R/ en français', *Actes du 4ème Congrès français d'Acoustique*, Marseille, Vol. 1, 417-420.

[Del69] Delattre, P.C. (1969) ' L' [R] parisien et autres sons du pharynx', *The French Review*, Vol. XLIII, 5-22.

[Mra88] Mrayati, M., Carré, R. and Guérin (1988) 'Distinctive regions and modes': A new theory of speech production', *Speech Communication*, Vol. 7, 257-286.

[Öhm66] Öhman, S.E.G. (1966) ' Coarticulation in VCV Utterances : Spectrographic Measurements, J. Acoust. Soc. Am. Vol. 39, 151-168.

[Öhm97] Öhman, S.E.G. (1997)'Coarticulation and the so-called Invariance Problem', In 'Polyphonie pour Ivan Fonagy, 359-366, l'Harmattan , Paris, Montréal.

[Roc71] Rochette, C. (1971) *'Les groupes de consonnes en français : Etude de l'enchaînement articulatoire à l'aide de la radiocinématographie'*, Thèse de doctorat d'état, Université de Strasbourg, 603 pages.

[San72] Santerre, Laurent. (1972)'Transitions articulatoires et transitions acoustiques dans la parole réelle', *3èmes Journées d'étude sur la parole*, Lannion,Vol.1,49-101,GALF.

Analyse phonétique des séquences [tʃ] et [dʒ] en français

Caroline Corneau

Laboratoire de Phonologie, Université Libre de Bruxelles
50, av. F.-D. Roosevelt, CP 175, 1050 Bruxelles - Belgique
Tél. :+32 2 650 45 07 - Fax :+32 2 650 24 50 - email :ddemoli@ulb.ac.be

ABSTRACT

This study presents a phonetic analysis of [tʃ] and [dʒ] sequences produced in short French sentences by four native speakers of French. This analysis is primarily based on acoustic and electropalatographic data. The durations of the 4 consonants as produced in the two sequences are given in order to establish a general temporal order pattern. Differences between acoustic and EPG data about closure duration are also examined. The following contact patterns were selected for discussion : first closure frame, maximal closure frame, release frame and most stable constriction frame. Subjects'articulatory habits as well as a devoicing phenomenon present in all four subjects are also examined.

1. INTRODUCTION

Cette étude présente une analyse phonétique des séquences occlusive dentale et fricative palatale sourde [tʃ] et sonore [dʒ] produites en parole continue par quatre sujets en français. En français, ces séquences sont considérées du point de vue phonologique comme des séquences de deux consonnes distinctes et non comme des affriquées monophonématiques [Cal89]. Le but de cette étude est de décrire les caractéristiques phonétiques de ces séquences afin d'en extraire, à travers la production des quatre sujets et les différents environnements vocaliques, certains des traits généraux qui les constituent. Le but ultime sera de comparer les séquences [t+ʃ] et [d+ʒ] avec le phénomène de palatalisation en français ainsi qu'avec les affriquées de l'anglais, afin d'insérer ces résultats dans la perspective de la phonologie articulatoire. Les paramètres qui seront examinés sont le signal acoustique, la durée des deux parties des séquences de consonnes, ainsi que la description articulatoire fournie par les données électropalatographiques. Cette dernière sera comparée au signal acoustique et aux mesures de durée. Les phénomènes mis en évidence lors de l'analyse seront discutés dans l'optique d'établir des hypothèses quant aux stratégies employées par les sujets lors de la production de ces séquences.

2. MATERIEL ET METHODE

L'ensemble des paramètres de l'expérience furent acquis via la station de travail Physiologia [Tes83], qui permet d'acquérir et de traiter simultanément le signal acoustique, les paramètres aérodynamiques, l'EPG et l'EGG. L'éléctropalatographe (EPG) utilisé est le système EPG2 de Reading. Le système utilise un palais artificiel en acrylique de 0,8mm d'épaisseur et contenant 62 électrodes de contact [Har89]. L'éléctroglottogramme utilisé est le système de Fourcin. Les paramètres aérodynamiques mesurés sont le pression intra-orale (PIO) et le débit d'air buccal (DAB). La PIO est mesurée avec un tube flexible en caoutchouc (2mm de diamètre interne) introduit dans la partie supérieure de l'oro-pharynx via la cavité nasale. Le DAB est mesuré avec un masque flexible en silicone. Le corpus se compose de 21 phrases :10 d'entre elles contiennent la séquence sourde [tʃ], et 11 la séquence sonore [dʒ] (figure 4). Ces séquences apparaissent toujours à l'intérieur d'un mot, et toujours dans le contexte -VCCV-. Elles apparaissent soit en début du mot (10 phrases pour [dʒ] et 5 phrases pour [tʃ]), soit à l'intérieur du mot (1 phrase pour [dʒ] et 5 phrases pour [tʃ]). Les voyelles qui précèdent la séquence sont assez variées et ne constituèrent pas, au départ, un critère pour le choix des mots à analyser. Les voyelles qui suivent la séquence sont quant à elles contrôlées de manière à avoir, pour chaque séquence, 2 ou 3 mots avec chacune des 5 voyelles choisies ([i-a-u-ɔ-e]).Ces 21 phrases furent produites une fois par 2 sujets féminins (CC et VL) et 2 sujets masculins (AS et DD), de façon totalement libre. Bien que la position de l'accent puisse avoir des répercussions sur l'articulation de ces séquences, son influence ne fait pas l'objet de cette étude et n'a donc pas été spécifiquement traitée. Les mesures temporelles effectuées sur les séquences combinent les informations électropalatographiques et acoustiques. La durée de l'occlusive correspond à l'intervalle de temps qui sépare la première occlusion totale (c'est-à-dire la première image EPG montrant une rangée horizontale complète d'électrodes activées) du relâchement de l'occlusion (c'est-à-dire l'image EPG suivant la dernière image qui présente une rangée horizontale complète d'électrodes activées). La durée de la fricative concerne l'intervalle temporel qui sépare ce second point du début de la voyelle (d'après le signal acoustique). Dans la plupart des cas, les mesures EPG correspondaient aux données acoustiques. Pour les phrases où ces deux données ne correspondent pas (voir § 4.1), la mesure de l'occlusive fut également prise sur base du signal acoustique. En ce qui concerne la configuration spatiale des contacts linguaux-palataux, différentes données EPG furent relevées : la

première occlusion, l'occlusion maximale, le relâchement de l'occlusion et la tenue la plus stable de la constriction. Il est à noter que la méthode décrite n'a pu s'appliquer à 2 items de CC et à 15 items de DD (voir § 4.1). Dans ces cas, nous avons seulement eu recours à des mesures acoustiques prises sur le signal de parole.

3. RESULTATS

3.1. CARACTERISTIQUES TEMPORELLES

La table 1 montre la durée, en ms, de chaque partie des deux séquences de consonnes, pour chaque sujet. La table 2 montre, par sujet et par séquence, le pourcentage d'items pour lesquels la partie occlusive de la séquence est respectivement inférieure, égale ou supérieure à la partie fricative.

Table 1

	t	ʃ	d	ʒ
AS	87,1	111,8	80,1	53,4
DD	86,7	93,7	82,5	59,9
CC	82,4	103,1	63,6	51,8
VL	130,7	114,3	123,8	58,7

Table 2

	Occl<fric	occl=fric	occl>fric
AS sonore	0%	9%	91%
AS sourde	70%	10%	20%
DD sonore	18%	27%	55%
DD sourde	50%	10%	40%
CC sonore	18%	18%	64%
CC sourde	70%	10%	20%
VL sonore	10%	0%	90%
VL sourde	30%	10%	60%

3.2. CONFIGURATION SPATIALE DES GESTES DE CONSTRICTION LINGUALE

Les données relevées sont examinées sous plusieurs aspects afin de caractériser les articulations des 4 sujets. Le lieu de l'articulation (pour l'occlusion maximale (figure 1) et pour la constriction (figure 2)) est décrit en termes du nombre moyen de contacts par rang, le premier rang étant le plus antérieur. Le nombre moyen de contacts établi aux 4 moments choisis est donné dans la table 3 pour [dʒ], et dans la table 4 pour [tʃ]. La figure 1 montre que l'articulation de CC et DD est plus antérieure que celle de AS et VL. Chez AS, la constriction laisse le plus souvent un espace central de 2 ou 3 électrodes non contactées sur le 4° rang et de 3 à 4 sur le 5° rang. Les rangs postérieurs ont généralement un "trou" de 4 électrodes au milieu. Chez DD, la constriction est asymétrique, décalée sur la gauche. L'ouverture est de 2 à 3 électrodes sur les 2° et 3° rangs, et de 2 à 4 sur le 4° rang. Les rangs postérieurs ont un trou plus central de 3 ou 4 électrodes. Pour CC, on a 2 à 3 trous sur le 2° rang, et 3 à 4 trous sur le 3° rang. Les rangs postérieurs sont plus ouverts : de 5 à 6 trous. Pour VL, on a 2 à 3 trous sur les 2°, 3° et 4° rangs. Les rangs postérieurs ont une ouverture de 4 à 5 trous.

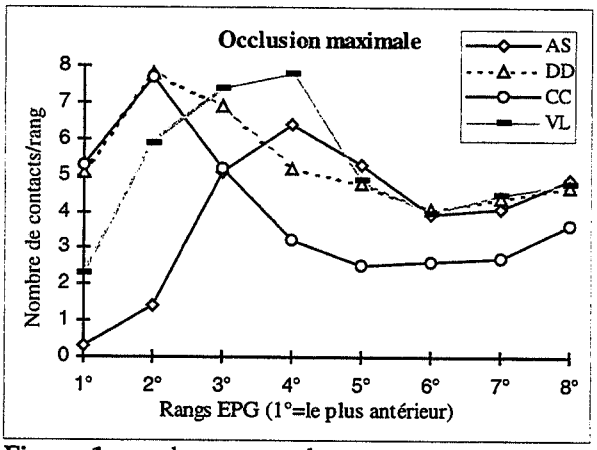


Figure 1: nombre moyen de contacts par rang pour l'occlusion maximale (pour les 2 séquences).

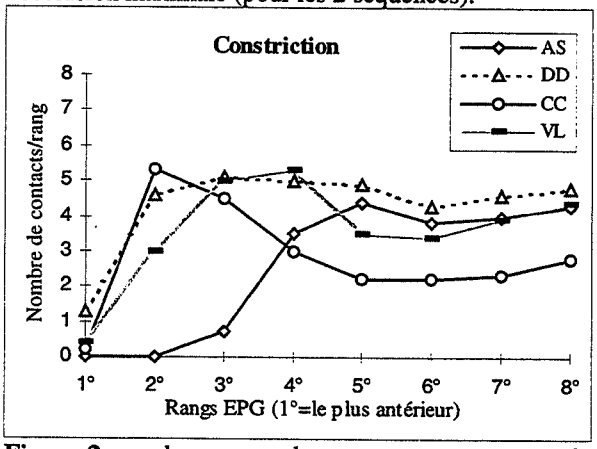


Figure 2: nombre moyen de contacts par rang pour la constriction la plus stable (pour les 2 séquences).

Table 3

	AS dʒ	DD dʒ	CC dʒ	VL dʒ
1° occl.	28,7	38,6	28,8	33,6
occl. max.	32,5	42,3	33,9	41,3
relâchement	27,1	37,4	24,6	33,9
constriction	21,7	35,7	22,7	29,3

Table 4

	AS tʃ	DD tʃ	CC tʃ	VL tʃ
1° occl.	20,8	35	25,3	28,3
occl. max.	30	42,9	32	40,9
relâchement	26,1	38,6	25,8	33,2
constriction	19,1	33,8	23,2	28,2

3.3. EGG

Certaines occurrences de [dʒ] présentent un dévoisement de la fricative. Ceci concerne 3 mots de

AS, 2 mots de DD mais surtout CC et VL, pour qui tous ces mots présentent le dévoisement (figure 3). La table 5 présente, en ms, les durées moyennes (pour 3 items de AS, CC et VL et 2 items de DD) de la partie voisée de l'occlusive, de la partie dévoisée de l'occlusive et de la fricative dévoisée.

Table 5

	AS	DD	CC	VL
occl. voisée	75,4	63,3	44,9	71,5
occl. dévois.	7,2	23,4	25,4	32
fric. dévoisée	67,8	81,1	45,6	43,6

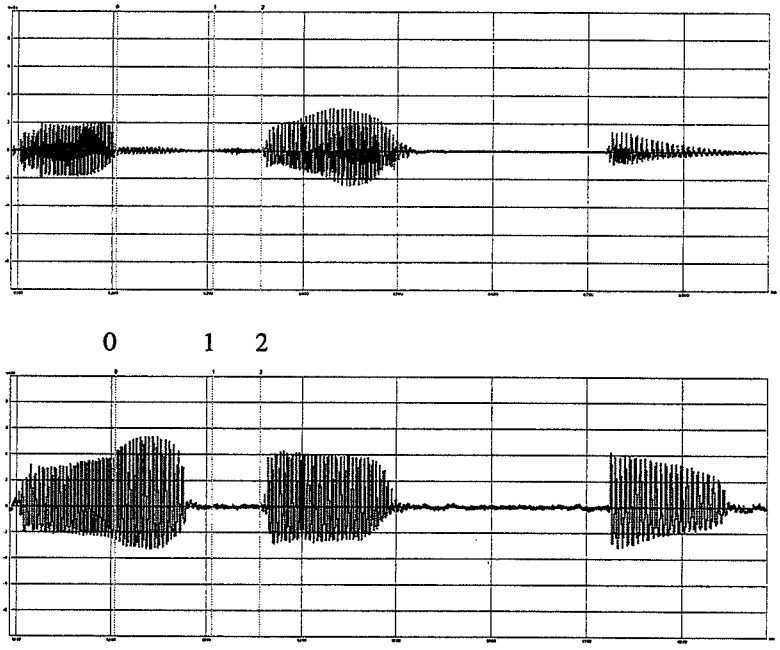


Figure 3:signal de parole (1) et EGG (2) pour le mot "adjacent" prononcé par le sujet VL. 0 indique le début de l'occlusion, 1 le relâchement de l'occlusion et 2 la fin de la fricative. L'EGG montre que la fin de l'occlusion et la fricative sont dévoisées.

4. DISCUSSION

4.1. OCCLUSIVE ET OCCLUSION

Si une occlusive est généralement caractérisée, au niveau articulatoire, par une occlusion totale sur une ou plusieurs rangées, ce n'est pas forcément toujours le cas. Ceci se remarque le plus chez DD, pour qui 5 [ʤ] ne montrent aucune occlusion totale. Ce même phénomène apparaît pour 2 [ʤ] de CC. Encore chez DD, 5 [ʧ] et 2 [ʤ] présentent une occlusion EPG nettement plus brève que celle mesurée sur le signal acoustique :le contact EPG complet apparaît après le début de l'occlusion acoustique et est relâchée avant l'apparition du bruit de friction. On note par exemple une occlusion EPG de 21.1ms qui apparaît 68.4ms après le début de l'occlusion acoustique et qui se relâche 19.3ms avant l'apparition du bruit de friction. De même, 1 [ʧ] et 2 [ʤ] de DD sont caractérisés par une occlusion se relâchant environ 10 ms avant l'apparition du bruit. Les données EPG ne sont donc pas toujours parfaitement synchronisées aux données acoustiques. Ceci rappelle les résultats de [Sho93] qui montrent que le passage par zéro de débit d'air buccal et l'occlusion EPG ne sont pas toujours simultanés. Une occlusive peut donc être sous-tendue par une articulation relâchée où subsiste une ouverture, qui est cependant minime: une électrode non contactée sur le(s) rang(s) avec le maximum de contacts. Cette constriction suffit à faire monter la pression intra-orale de manière comparable à une occlusive, et ne modifie pas la forme du signal acoustique. C'est dans ces cas que la mesure des durées fut prise sur base acoustique, le relâchement de l'occlusion au niveau articulatoire ne correspondant pas au début réel de la fricative. Néanmoins, on peut, sur base de l'EPG, remarquer que la durée moyenne des occlusions pour DD est de 66,36 ms pour la séquence [ʧ] et de 52,53 ms pour la séquence [ʤ].

4.2. DUREE

Trois des quatre sujets présentent un schéma assez stable quant aux durées des 4 phonèmes produits dans ces séquences. Pour la séquence [ʤ], l'occlusive est plus longue que la fricative, alors que c'est l'inverse pour la séquence [ʧ] (cf table 1). Ces résultats permettent d'établir le fait suivant. On aura, par ordre décroissant de durée:[ʃ]>[t]>[d]>[ʒ]. Ce schéma n'est pas respecté par VL pour qui [ʃ] est plus bref que [t] et

[d], et plus long que [ʒ]. Ceci paraît dû à une durée plus longue des occlusives que des fricatives observées chez VL, par rapport aux autres sujets. Notons que de manière prévisible [Cal89] pour les 4 sujets, la consonne sourde (occlusive ou fricative) est plus longue que son correspondant sonore. Ceci permet de constater que malgré la variabilité inter-sujets, certaines constantes paraissent se définir dans la production de ces séquences de consonnes.

4.3. EPG

Les occlusions analysées ici sont généralement situées sur plusieurs rangs, qui sont soit totalement contactés, soit qui constituent ensemble l'occlusion. On a vu que le lieu de l'articulation varie selon les sujets. Ceci témoigne du fait que les habitudes articulatoires sont assez stables et ce malgré les différents environnements vocaliques. Si l'on regarde les séquences dont le lieu d'articulation ne correspond pas au cas général décrit ci-dessus, on voit que les contextes vocaliques sont assez variés, pour l'occlusion comme pour la constriction. Les contextes retrouvés le plus fréquemment pour les cas d'articulation plus antérieure sont, pour l'occlusion et la constriction: [i-ɔ] et [i-e] (chez les 4 sujets), mais aussi [a-i] (AS, VL), [i-u] et [ɶ̃-e] (DD, VL). Les cas d'articulation plus postérieure sont : [ɑ̃-i] et [a-u] (chez 3 sujets), mais aussi [a-a] et [ɶ̃-a] (2 sujets). On ne retrouve donc pas, ici, les résultats obtenus par [Mai96], pour qui l'articulation de [t] et [tʃ] est plus antérieure dans un contexte [a-a] et [e-e] que dans un contexte [i-i] et [u-u]. Il est difficile, d'après les résultats présents, d'obtenir une caractérisation générale inter-sujets sur l'influence du type de voyelle précédant ou suivant la séquence. Le nombre de contacts semble plus influencé par le type de voyelle qui précède la séquence que par celle qui suit. En effet, le nombre moyen de contacts lors de l'occlusion maximale est le plus souvent dépassé lorsque la voyelle qui précède la séquence est fermée et antérieure. Typiquement, parmi les 5 items précédés de [i] ([i-ɔ], [i-u] et [i-e]), le nombre moyen de contacts est dépassé chez AS (les 5 items), chez DD, chez CC et chez VL (4 items). Il est donc à noter que les habitudes articulatoires sont assez stables chez un même sujet et varient plus nettement entre les sujets qu'en fonction de l'environnement vocalique. Toutefois, les mots du corpus n'étant pas réellement contrôlés en ce qui concerne la voyelle qui précède les séquences de consonnes, le problème reste ouvert.

4.4. EGG

Les cas de dévoisement montrent que la séquence sonore ne reste pas forcément voisée tout le long de sa production. L'hypothèse que l'on peut émettre est que pour pouvoir générer la turbulence caractéristique de la fricative, le débit d'air doit être assez important. La manière de l'augmenter est d'ouvrir la glotte, ce qui provoque le dévoisement. En outre, la durée de la fricative « sonore » (et de la séquence entière) reste inférieure à celle de [ʃ], ce qui permet donc une identification correcte.

5. CONCLUSION

Cette étude fournit quelques uns des traits généraux caractérisant les séquences du français [t+ʃ] et [d+ʒ] étudiées. Un schéma d'ordre des durées relatives des 4 consonnes produites dans ces séquences a pu être établi. Certaines variations apparaissent néanmoins entre les sujets, tant dans l'aspect durée que dans l'aspect lieu d'articulation. Ces variations semblent selon cette étude relever davantage des habitudes articulatoires des sujets que du contexte vocalique, lesquelles semblent avoir une influence relativement faible sur le lieu de l'occlusion. Le nombre de contacts, quant à lui, paraît plus clairement dépendre du type de voyelle qui précède la séquence. Cette étude ne mettant pas en évidence une relation nette entre contexte vocalique et lieu d'articulation, cette question reste ouverte et devrait profiter de recherches plus ciblées sur ce problème.

REFERENCES

[Cal89] Calliope (1989), La parole et son traitement automatique, Masson.

[Sho93] Shockey, L. et Gibbon, F.(1993) « Stopless stops in connected English », *Speech Research Laboratory, University of Reading Work in Progress.*

[Mai96] Mair, S., Scully, C. et Shadle, C.(1996) « Distinctions between [t] and [tʃ] using electropalatography data ».*Proceedings of the fourth International Conference on Spoken Language Processing. Vol 5. Philadelphia.*

[Har89] Hardcastle, W., Jones, W., Knight, C., Trugdeon, A. and Calder, G. (1989), « New developments in EPG : a state of the art report », Clinical linguistics and Phonetics , 1 : 1-38.

[Tes83] Teston B. (1983), « Un système de mesure des paramètres aérodynamiques de la parole : le polyphonomètre modèle III », Travaux de l'Institut de Phonétique d'Aix. Vol. 9, pp. 373-383.

Figure 4

Mets ton jean bleu. Je suis allé à Djibouti. J'ai mangé un chicken filet. Voici monsieur Rothschild. Une pizza cheesy crust. Voici John Wayne. Nom di djos. Voici monsieur Tchombe. J'ai mangé des natchos grandes. Nom di djou. As-tu vu Jumanji ? L'enfant a fait atchoum. Un parfum au patchouli. Le train fait tchouk tchouk. Voici monsieur Jakobson. Un morceau de jazz. Prononcez le mot adjacent. Une pièce de Tchaïkovsky. Des scraboutchas. Voici James Ensor. Voici Jason.

Etude acoustique préliminaire des 16 voyelles orales du Paez de Talaga, langue amérindienne

Egidio MARSICO

DDL
MRASH
14, Avenue Berthelot
69363 Lyon cedex 07
France
Tél : 04-72-72-64-62 / Fax :04-72-72-65-90
Egidio.Marsico@mrash.fr

Tulio ROJAS CURIEUX

CCELA-Universidad de los Andes
Carrera la-este N°18 A10
Edificio P-Segundo piso
Apartado Aéreo 4976
Santafé de Bogotá-Colombia
Tél : 2824066Ext2571 / Fax : 2841890-2815771
trojas@uniandes.edu.co

Abstract

The vocalic system of Paez is typologicaly rare, less than 5% of the world's languages has more than 17 phonemic vowels and only 8,5% of them distinguishes four different qualities. It appears with acoustic analysis, that these qualities slightly shift toward more open ones when lengthened, and that the glottalized and aspirated series show a temporal organization in two different phases.

1. Introduction

1.1. Le système vocalique du Paez

On trouve dans la base de données UPSID (UCLA Phonological Segment Inventory Database, version 351 langues) une description du système phonologique du Paez (langue sud-amérindienne) faite par F. Gerdel [Ger73] et proposant le système vocalique suivant :

/i e a u/ ainsi que leur pendant nasalisé.

Des investigations plus récentes menées au sein du CCELA (Centro Colombiano de Estudios en Lenguas Aborigenes) de l'Université des Andes (Bogota, Colombie), sur quatre variétés dialectales du Paez, proposent des systèmes vocaliques beaucoup plus riches comportant parfois jusqu'à 32 voyelles dont des séries longues, glottalisées et aspirées.

Cette étude se propose d'apporter des mesures acoustiques sur l'un des parlers Paez.

1.2. Le corpus

Nous disposons d'enregistrements d'Abelardo RAMOS, locuteur natif et actif de la variante de Talaga, par ailleurs linguiste impliqué dans les recherches du CCELA (qu'il soit ici remercié). La collecte des données a été effectuée par Tulio Rojas dans le cadre d'une étude générale de la langue Paez [Roj91] et [Roj96]. Ceci explique que nous ne disposons que d'un nombre faible de répétitions et que par conséquent cette étude ne peut-être considérée que comme un travail préliminaire. Nous pensons toutefois que ces premières données acoustiques sur un système vocalique très original d'un point de vue typologique sont tout à fait intéressantes.

1. Le PAEZ

1.1. La localisation

Les Paez sont localisés en Colombie dans les départements du Caqueta, du Cauca, du Huila, du Putumayo, du Meta, du Tolima et du Valle del Cauca. Le département du Cauca est celui qui abrite la plupart des Paez (plus de 100.000 locuteurs). On trouve dans les autres départements, des regroupements plus ou moins importants qui sont le résultat de processus migratoires depuis leurs terres ancestrales. Le département du Cauca (Figure 1), se situe dans la région Sud-Ouest de la République de Colombie, sa topographie inclut des collines, des paramos (hauts plateaux andins) et des montagnes aux neiges éternelles. L'altitude de cette région est comprise entre 500 et plus de 4000 mètres. La température va de 25° C dans les zones basses à des valeurs négatives dans les zones les plus hautes.

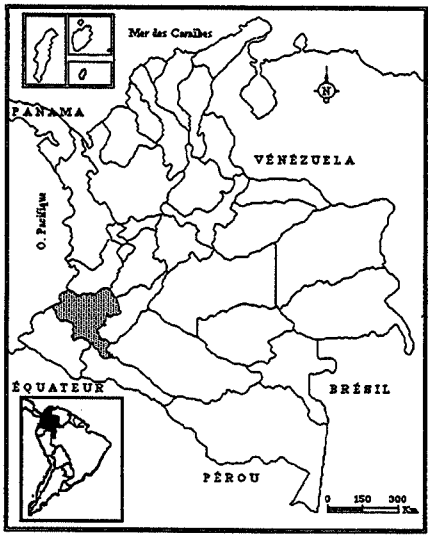


Figure1 : Localisation du peuple Paez en Colombie

1.2. Les variantes dialectales

La langue Paez (*nasa yuwe*, littéralement nasa signifie 'gens', 'être humain' et yuwe, 'bouche', 'mot', 'langue') a été inclue dans la famille Chibcha (cf. Ruhlen [Ruh87] mais une étude plus récente de Constenla [Con93], tend à l'en exclure, nous suivrons quant à nous la classification de Landaburu [LanPrép] qui donne une lignée unique du Paez.

A propos de la distribution dialectale, Rocío Nieves [Nie95] propose : "La description du Nasa Yuwe est en cours, c'est pourquoi nous n'avons pas de données permettant d'établir de véritables isoglosses qui mèneraient à une distribution précise des variétés régionales. Cependant, nous pourrions, à partir des études existantes et surtout des études phonologiques, proposer une première hypothèse de 'zonification'.

Nous aurions alors comme grandes divisions : une zone au nord regroupant Munchique-Tigres, Huellas, Toribio, San Francisco et Tacueyo. Au centre probablement deux zones, l'une formée par Jambalo et Pitayo, et l'autre par Caldono, Pioya et Pueblo Nuevo. Au sud il y aurait une zone avec Novirao et Paniquita. Enfin, nous pourrions penser à Tierradentro comme à une grande zone aux multiples distinctions donnant lieu à des subdivisions. Mais, répétons-le, ceci est une première approche [...] et seule une étude détaillée [...] permettrait de délimiter avec précisions les isoglosses." (notre traduction)

Nous donnons table 1 l'étude comparée de Nieves [Nie91] pour les systèmes vocaliques de trois variétés.

Table 1 : Systèmes vocaliques de 3 dialectes Paez

Voyelles	Munchique-Tigre	Toribío	Caldono
Orales	i e a u	i e a u	i e a u
Nasales	ĩ ẽ ã ũ	ĩ ẽ ã ũ	ĩ ẽ ã ũ
Longues			iː eː aː uː
Glottalisés	iʔ eʔ aʔ uʔ	iʔ eʔ aʔ uʔ	iʔ eʔ aʔ uʔ
Aspirées	i^h e^h a^h u^h	i^h e^h a^h u^h	i^h e^h a^h u^h

Nous la citons : "L'absence de voyelles longues en Munchique-Tigre et en Toribio saute aux yeux. Pour le reste, il n'y a pas de différences : dans toutes les régions il existe l'opposition entre orales et nasales, de même pour les corrélations d'interruption glottale et d'aspiration. Dans toutes les régions, on peut également combiner les traits des différentes oppositions, on obtient ainsi des voyelles brèves orales, brèves nasales, glottalisées orales, glottalisées nasales, aspirées orales et aspirées nasales".

Notons pour compléter cette citation qu'aucun système combinant sur le même timbre glottalisation et aspiration n'a été rencontré, ces traits semblent donc mutuellement exclusifs.

1.3. La variante de Talaga

Cette variante se rencontre dans la zone de Tierradentro, son système vocalique est en fait similaire à celui de la variante de Caldono, d'après Nieves [Nie91], ce serait la variante au système le plus stable. Son système vocalique consiste en fait en une base de 4 segments déclinés en 8 sous-systèmes pour lesquels nous proposons la schématisation illustrée Figure 2.

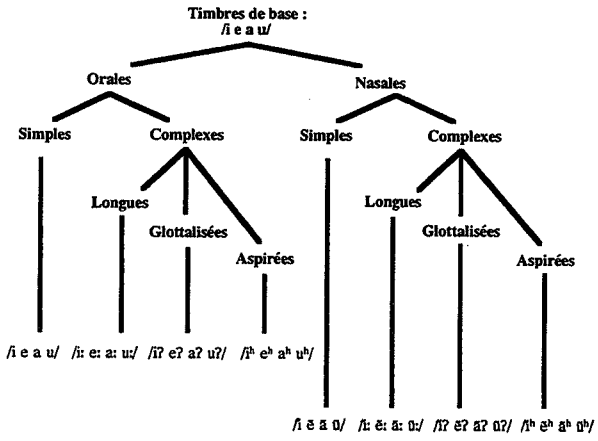


Figure 2 : Système vocalique du Paez de Talaga

2. L'analyse acoustique

Nous examinerons ici les 16 voyelles orales, selon le corpus suivant :

Orale brève :		
/i/	isa	'veritable'
/e/	use	's'approcher du chaud'
/a/	isa	'veritable'
/u/	use	's'approcher du chaud'
Orale longues:		
/iː/	kiːs	'lever'
/eː/	neːs	'formation'
/aː/	maːn	'lequel est-ce?'
/uː/	kuːs	'au petit matin'
Orales glottalisées :		
/iʔ/	kiʔs	'régime'
/eʔ/	peʔs	'soeur'
/aʔ/	aʔte	'lune'
/uʔ/	kuʔta	'bras'
Orales aspirées :		
/i^h/	tʃi^hme	'blanc'
/e^h/	tse^{h}me	'visage mal rasé'
/a^h/	a^h	'fumée'
/u^h/	tu^{h}me	'gris-argent'

La digitalisation à partir d'une cassette analogique a été faite avec le logiciel SoundForge en 22Khz, 16 bits, mono. Les analyses du signal ont été réalisées sur Multispeech.

2.1. Le protocole

Nous disposions de trois répétitions pour chaque mot, et pour chaque répétition nous avons pris trois mesures : au centre de la voyelle, à +15ms et à -15ms par rapport au centre.

Pour les séries glottalisée et aspirée, comme signal et spectrogramme s'accordaient à présenter deux parties clairement distinctes, nous avons pris systématiquement deux séries de trois mesures.

Avant de passer aux commentaires des quatre ensembles de voyelles, nous tenons à rappeler une nouvelle fois le caractère préliminaire de cette étude qui bien que ne pouvant présenter de mesures statistiques étant donnée le faible nombre de nos mesures (1 locuteur, 3 répétitions), s'attachera néanmoins à décrire ce système vocalique très original.

2.2. *Les timbres de bases*

Nous donnons Table 2 les valeurs moyennes de F1 et de F2 pour les quatre timbres vocaliques:

Table 2 : F1 et F2 moyens des orales brèves

	i	e	a	u
F1	279	442	564	280
F2	2595	2126	1730	549

Ceci révèle une architecture en 3 niveaux d'aperture, avec une non-équidistance entre i-e et e-a. Il convient de remarquer que les systèmes à 4 timbres sont typologiquement rares, puisque d'après Maddieson [Mad84], seulement 8,5 % des langues présenteraient cette organisation.

2.3. *La série longue*

Hormis pour le /a/, les trois autres voyelles voient leur F1 considérablement augmenter (Table 3), c'est à dire que l'articulation de i, e, u est plus ouverte quand elle est prolongée. Notons également la 'réorganisation horizontale' de /e/ et /a/ que révèlent les variations de leur F2.

Table 3 : F1 et F2 moyens des orales longues

	i:	e:	a:	u:
F1	356	546	557	401
F2	2561	2288	1609	655

2.4. *La série glottalisée*

La première particularité de ces voyelles est le fait que l'on distingue clairement deux parties différentes dans le signal. Ces deux zones consécutives sont de longueur inégale et le ratio de la première sur la deuxième varie en fonction des voyelles. Nous appellerons zone A la première partie du signal qui correspond en fait au timbre tel qu'il serait prononcé sans trait secondaire (hormis les quelques variations que nous commenterons plus tard), et zone B la deuxième, qui est en fait la phase glottalisée. La Table 4 donne les durées moyennes relevées en ms :

Table 4 : Longueurs relatives des zones A et B

	i:	e:	a:	u:
zone A	105	66	103	77
zone B	79	105	67	68
total V	184	171	170	145
A/B	1,33	0,63	1,54	1,13

Il est intéressant de remarquer l'inversion du ratio A/B pour le /e/, bien que nous ne soyons pas, pour le moment, en mesure de l'expliquer.

La figure 3 donne les analyses du /e?/ dans le mot /pe?s/, on y trouve un spectrogramme en bande large et un en bande étroite (respectivement fenêtres B et C, la courbe d'amplitude (D), le pitch (E) et le signal en A.

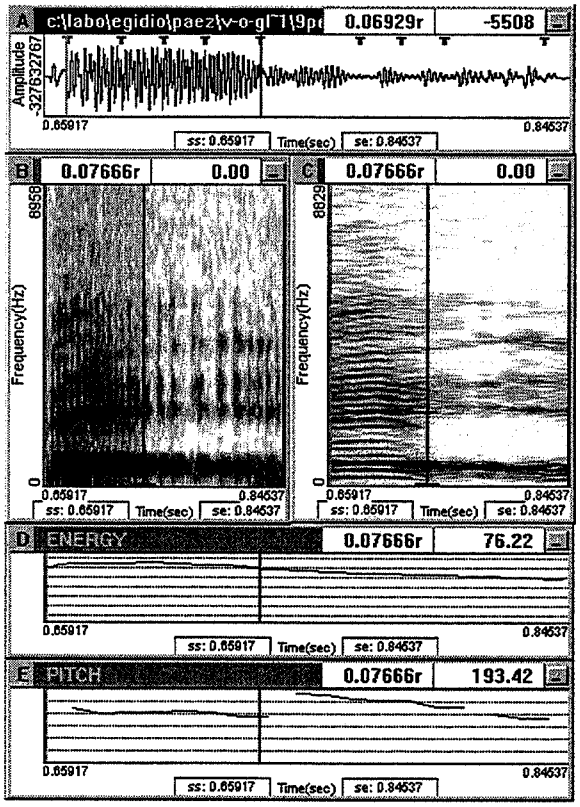


Figure 3 : Analyses acoustiques du /e?/ de /pe?s/

Nous pouvons signaler à propos de la zone A, que parallèlement à ce qui se passe avec le F1 des voyelles longues, celui-ci augmente davantage encore (Table 5) et même celui du /a/ gagne environ 40 Hz.

Table 5 : F1 et F2 moyens des glottalisées (zone A)

	i?	e?	a?	u?
F1	399	551	603	419
F2	2578	2153	1406	674

La zone B quant à elle offre deux particularités. La première est ce changement flagrant observé dans la fréquence de vibration des cordes vocales, et la deuxième une baisse importante de l'amplitude du signal. Ceci nous amène à citer Ladefoged et Maddieson [Lad95] à propos des voyelles 'creaky' ou craquées : "En règle générale, les voyelles articulées avec une 'voix craquée' ont plus d'énergie dans les harmoniques de la région du premier et du second formants que celles à voix modale. [...] Il y a aussi une tendance (bien que non présente dans toutes les langues) pour les voyelles 'craquées' à avoir une fréquence de vibration des cordes vocales plus irrégulière." Certes nous devons relativiser cette remarque. Tout d'abord en précisant qu'il s'agit ici de

variation intra-voyelle et non d'un changement de phonation portant sur toute la longueur du segment (qui s'opposerait à un timbre basique). D'autre part, le spectrogramme à bandes étroites fait état non pas d'un renforcement d'énergie dans les zones du premier et du second formants, mais en quelques sortes d'une 'conservation' de cette dernière eu égard à la désertion d'énergie observée à l'entour. Ceci fait que l'on retrouve des valeurs de F1 et F2 quasi inchangées par rapport aux valeurs mesurées dans la zone A. Les variations de F0 sont trop irrégulières en fonction de la voyelle pour que nous puissions en proposer une interprétation.

2.5. La série aspirée

Précisons dès l'abord que la difficulté potentielle à segmenter ce type de voyelle a pu être contournée par le fait que ces segments étaient -hormis le /a^h/- tous inclus dans un mot. La figure 4 montre l'exemple du /e^h/ de /tse^{h}me/.

On remarque évidemment toujours cette séparation en deux zones, la première demeurant l'équivalent d'un timbre basique. Cependant les variations de F1 par rapport aux pendants orals brefs sont moins uniformes, en effet, la valeur moyenne demeure constante pour /i/, elle baisse de 90Hz pour /a/, et augmente de 80Hz pour /e/ et /u/. La zone B attire également d'autres commentaires que nous pourrions résumer en disant qu'elle semble correspondre à une 'extinction' progressive de la voyelle.

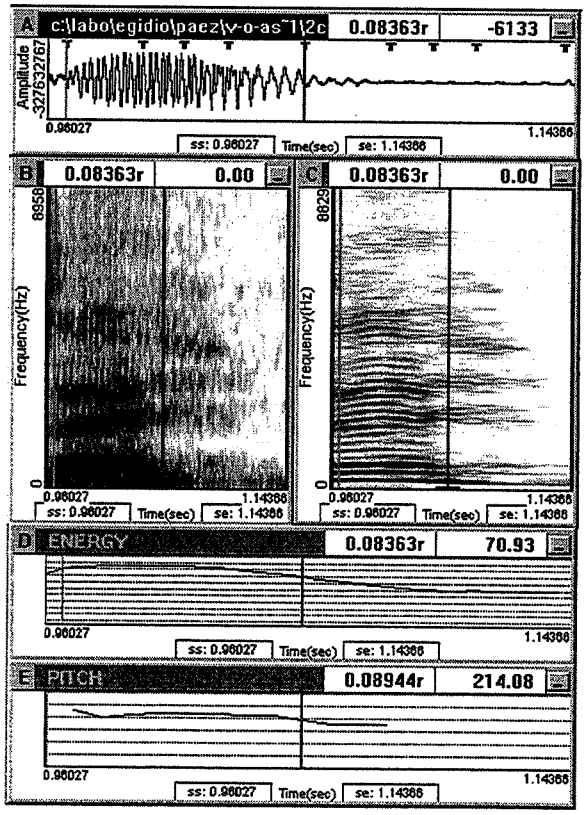


Figure 4 : Analyses acoustique du /e^h/ de /tse^{h}me/

3. Conclusion

Le système vocalique du Paez de Talaga est donc un système 'à part'. D'un point de vue typologique, seulement 4,1% des langues ont plus de 17 phonèmes vocaliques (il en compte 32) et 8,5% ont une base à 4 timbres. Notons que la quatrième, /e/, est antérieure. L'analyse acoustique révèle un changement de timbres coincidant avec l'allongement, et une organisation temporelle des voyelles glottalisées et aspirées en deux phases.

4.Bibliographie

[Con93] Constenla A. (1993), "La familia Chibcha", Estado Actual de la clasificatión de las lenguas Indígenas de Colombia, Ponencias presentadas en el Seminario-Taller realizado en el Instituto Caro y Cuervo, publicado en Biblioteca "Ezequiel Uricoechea" núm 11. Santafé de Bogotá : Imprenta Patriótica del Instituto Caro y Cuervo, 1993 (75-125).

[Ger73] Gerdel F. (1973), 'Paez phonemics', Linguistics 104 : 28-48.

[Lad95] Ladefoged P. & Maddieson I. (1995), Sounds of The World's Languages, Blackwell Publishers.

[LanPrép] Landaburu J. "Clasificación de las lenguas Indígenas de Colombia", Lenguas Indígenas de Colombia. Estudio Preliminar para un Atlas Lingüístico. Santafé de Bogotá : Imprenta Patriótica del Instituto Caro y Cuervo, (en prensa).

[Mad84] Maddieson I. (1984), Patterns of Sounds, Cambridge University Press, Cambridge.

[Nie95] Nieves R. (1995), "Hacia una zonificación dialectal de la lengua páez", Colombia País Plurilingüe. Boletín de Lingüística Amerindia y Afroamericana N° 3. Cali : Universidad del Valle. (10-18).

[Nie91]Nieves, R. (1991), "Fonología comparada" in Estudios Fonológicos de la Lengua Páez (nasa yuwe) . Volumen 6, serie Descripciones. Bogotá : CCELA-Universidad de Los Andes. (121-149).

[Roj96] Rojas Curieux T. (1996) Prédication, aspect et modalité dans la langue paez - langue amerindienne de Colombie. Thèse de troisième cycle. Université de Paris 7. (498 pp).

[Roj91] Rojas Curieux T. (1991), "Fonología variante munchique" en Nieves. (ed), - Estudios Fonológicos de la Lengua Páez (nasa yuwe). Volumen 6, serie Descripciones. Bogotá : CCELA- Universidad de Los Andes. 1991 (43-95).

[Ruh87] Ruhlen M. (1987), Classification of the world's languages, vol. 1 : Classification, Stanford University Press, Stanford California.

Les systèmes consonantiques
Des tendances universelles à l'ontogenèse

Nathalie VALLEE , Louis-Jean BOË, Muriel STEFANUTO

Institut de la Communication Parlée, INPG/Université Stendhal
BP 25X 38040 Grenoble cedex
Tél. : 04 76 51 76 45 - Fax : 04 76 82 43 35
vallee@icp.inpg.fr

ABSTRACT

Within the frame of a substance-based linguistic approach, we present here typological researches on consonant systems from UPSID (*UCLA Phonological Segment Inventory Database*, [Mad86] [Mad89]. We implemented UPSID at ICP and we have analysed it with the help of an original methodology. We show a set of general orientations which not only confirm or make more precise some regularities already stressed on, but also brings up new data. The search for universal tendencies in consonant systems of languages of the world is a necessary step for a theoretical perspective in syllable structure. Finally we discuss the general tendencies in the frame of ontogenesis.

1. INTRODUCTION

Depuis Troubetzkoy et ses *Principes* (1939), la taxinomie n'est plus seulement une démarche de la linguistique historique : associée à la recherche d'universaux en synchronie elle constitue aujourd'hui une étape importante dans l'élaboration des théories linguistiques. Institutionnalisé en 1961, sous l'égide du *Social Science Research Council*, lors de la *Conference on Language Universals* de New-York, ce champ de recherche a pour objectif la mise en évidence des structures de base communes à toutes les langues en synchronie comme en diachronie. *The Language Universal Project* (1967-1976) a permis la constitution des *Stanford Phonology Archives* [Gre78] à partir desquelles sont nés des travaux d'importance concernant les classifications typologiques et l'étude des universaux phonologiques de [Sed69] [Cro78] à [Mad86] [Val94]. Mais toutes ces études, il faut le préciser, varient dans leur contenu : les données sont sans cesse complétées et améliorées, et les interrogations sur le matériau diffèrent selon les auteurs. Depuis Troubetzkoy, les travaux typologiques ont davantage porté sur les systèmes vocaliques. Cet état de fait est probablement à relier au nombre plus important de consonnes dans les systèmes et à la multiplicité des paramètres de classement, rendant plus ardu l'émergence des types. Citons parmi les taxinomies consonantiques les travaux de [Hoc55] [Hag82] [Mad86] [Lin88] [Lav94] [Ste96].

Constituées à partir des archives de *Stanford*, les bases UPSID (UCLA Phonological Segment Inventory Database), d'abord 317 [Mad86] puis 451 langues [Mad89], contiennent des données phonologiques génétiquement pondérées et descriptivement harmonisées ([Mad86], 5–6). Nous les avons implantées à l'ICP pour nourrir les recherches typologiques sur les voyelles, les diphtongues et les consonnes.

Nous présentons ici un ensemble de tendances des systèmes consonantiques que nous rapprochons de données disponibles sur l'acquisition. En effet, parmi les différentes étapes de l'ontogenèse, celle du babillage apparaît comme fondamentale puisqu'elle marque les « précurseurs » des attributs universels du langage, au sens défini par [Mac90].

2. TENDANCES UNIVERSELLES

2.1 Éléments de taxinomie

La classification traditionnelle des consonnes repose sur les lieux et les modes d'articulation. Ces paramètres constituent bien évidemment les dimensions élémentaires de notre étude typologique.

UPSID451 répertorie 920 phonèmes dont 654 segments consonantiques, répartis sur 13 lieux d'articulation (Figure 1) et une quinzaine de modes. On dénombre 153 plosives, 61 implosives, occlusives éjectives et clicks non affriqués, 134 fricatives, fricatives éjectives, 155 affriquées, affriquées éjectives et clicks affriqués, 95 vibrantes, battues et approximantes centrales et latérales, 51 nasales et 5 consonnes de type *h sounds*.

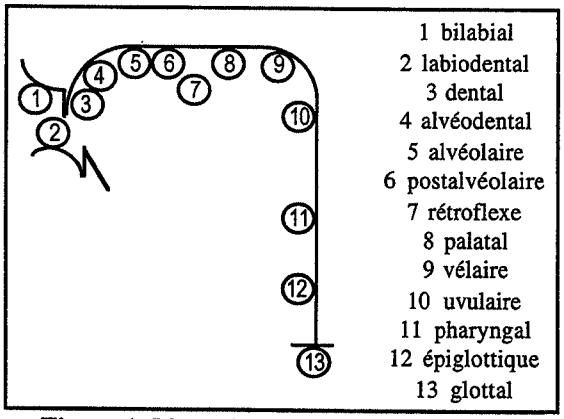


Figure 1. Lieux d'articulation consonantiques (UPSID451).

Les modes ont été regroupés sous 7 catégories classant ainsi les consonnes en plosives (dont implosives, éjectives, occlusives glottales), nasales, fricatives (dont celles éjectives et *h sounds*), affriquées (dont éjectives), approximantes, vibrantes/battues, clicks.

Nous avons élaboré notre typologie avec une détermination des lieux plus détaillée [Mad86] en composant avec celles proposées dans [Cre94] et [Lad96]. Bien que désignée par un mode, nous avons conservé la classe des rétroflexes comme le proposent [Mad86], [Lad96] et l'API de 1996. Une répartition de cette catégorie de consonnes, qui recouvrent plusieurs lieux (alvéodental, alvéolaire, prépalatal...), nécessiterait une investigation des sources. Soulignons que la comparaison avec les taxinomies antérieures n'est toutefois pas aisée car toutes diffèrent dans la préparation des données et les méthodes de classement.

2.2 *Distributions*

Les systèmes phonologiques ont le plus souvent entre 18 et 25 consonnes, avec un mode à 22 : minimum 6 pour le rotokas (langue indo-pacifique, 11 phonèmes) et maximum 95 (dont 48 clicks) pour le !xũ (famille khoisan, 141 phonèmes). La moyenne par langue est de 22 consonnes dont 7.8 plosives, 4.1 fricatives, 3.3 nasales, 2.9 approximantes, 2 affriquées, 0.6 éjective, 0.52 vibrante ou battue, 0.25 implosive, 0.2 click.

La Table 1 donne la répartition des consonnes (654 au total) par lieux d'articulation. Les plus fréquentes, tous modes confondus, sont les alvéodentales (15.3%), suivies des bilabiales (14.3%) puis des vélaires (12.6%). Le regroupement des consonnes dentales, alvéodentales, alvéolaires, postalvéolaires et de l'ensemble des rétroflexes en une catégorie *coronale* permet de mieux apprécier l'écrasante proportion de celle-ci : 44.5% dans $UPSID_{317}$. Bien que cette catégorie regroupe une grande diversité d'articulations [Lad96], ce choix est tout à fait acceptable pour l'élaboration de notre typologie. En effet, le trait de lieu ne sert que rarement à distinguer phonologiquement des consonnes de ce type, sauf pour les affriquées, et en tout cas pas entre coronales antérieures [Kea90].

Table 1. Fréquences d'occurrences des lieux d'articulation consonantiques ($UPSID_{317}$).

alvéodental	15.3%
bilabial	14.3%
vélaire	12.6%
palatal	10%
apico-alvéolaire	9.9%
glottal	7.6%
lamino-postalvéolaire	6.7%
labiovélaire	5.1%
dental	5%
lamino-alvéolaire	3.4%
labiodental	3.3%
rétroflexe	2.9%
uvulaire	1.7%
autres	<1%

La répartition des consonnes d'$UPSID_{317}$ par mode montre que certains sont plus exploités que d'autres : plosives 38.6%, fricatives 20.2%, nasales 14.6%, approximantes 13%, affriquées 9.6%, vibrantes ou battues 3.9%. Parmi les plus fréquentes arrivent en tête les occlusives coronales (type /t/), orales et sourdes (97.5%), les nasales bilabiales et coronales (types /m/ et /n/, présentes aussi dans plus de 9 langues sur 10), suivies de /k/, /j/, /p/ rencontrées dans plus de 80% des langues, /w/ et /s/ dans 2 langues sur 3, /d/, /b/, /h/, dans plus de 60%, /l/, /g/, /ŋ/, /ʃ/ dans la moitié des langues.

2.3 *Systèmes*

2.3.1 *Taille des systèmes de lieux et modes d'articulation*

L'élaboration d'une typologie par lieux et modes des systèmes consonantiques fait apparaître une forte corrélation entre la taille des systèmes de lieux et les différents modes d'articulation (Table 2).

- Dans la plupart des langues, les plosives et les nasales mettent en jeu respectivement de 3 à 6, et de 1 à 6 lieux. Le mode plosif est le seul universellement exploité par les systèmes sonores. Il est aussi le seul à utiliser systématiquement les oppositions de lieux : les langues en distinguent au moins trois.
- Les fricatives et les approximantes constituent les catégories consonantiques qui recrutent le plus grand nombre d'oppositions de lieux : le mode fricatif est beaucoup plus largement distribué entre les différentes tailles de systèmes. Si le jeu des oppositions de lieux des fricatives est souvent plus complexe que celui des plosives, il est important de préciser qu'il n'est pas exploité par 8% des langues de l'échantillon – alors que toutes possèdent des plosives.
- Dans la catégorie des *trills, flaps* et *taps* (vibrantes et battues), les systèmes ne présentent que rarement une opposition de lieu, et en aucun cas ils n'en distinguent plus de 2. Cette même tendance se retrouve pour les affriquées.

Plus généralement :

- Les langues exploitent systématiquement l'opposition de lieu pour les consonnes plosives ; près de la moitié les répartissent dans des systèmes à 4 lieux.
- L'opposition de lieu est rare pour les affriquées, exceptionnelle pour les vibrantes/battues.
- Les approximantes sont de préférence réparties sur 3 lieux d'articulation.
- Les fricatives se répartissent plus largement dans diverses tailles de systèmes de lieux.

2.3.2 *Systèmes de lieux par modes*

La Table 3 rend compte du contenu des systèmes de lieux par mode d'articulation. Sont représentées les répartitions des plosives, nasales, fricatives et affriquées par nombre de lieux d'articulation.

Il n'existe pas de langues sans plosives, ni même n'en ayant qu'une seule. Lorsqu'un système les répartit sur 2 lieux, l'un est coronal, l'autre bilabial ; sur 3 lieux : coronal, bilabial et vélaire. C'est à 4 qu'apparaît le lieu glottal ; si le système répartit ses consonnes entre 5

Table 2. Taille des systèmes de lieux et leur représentativité dans $UPSID_{317}$.
Le nombre d'oppositions de lieux varie selon les modes ; de 0 (c'est-à-dire 1 lieu par système) à 28 (8 lieux). Les valeurs présentées ne tiennent pas compte du trait de voisement : 28.4% des 317 langues ont un système de plosives à 3 lieux, 43.2% possèdent des plosives réparties sur 4 lieux, etc.

	1 lieu	2 lieux	3 lieux	4 lieux	5 lieux	6 lieux	7 lieux	8 lieux
trills, flaps, taps	67.2%	4.7%						
affriquées	44.8%	20.8%	6.9%	0.3%				
plosives			28.4%	43.2%	22.4%	5.9%		
nasales	2.5%	31.5%	30.6%	25.9%	4.5%	2.2%		
approximantes	8.2%	26.5%	49.8%	8.2%	2.5%	0.3%	0.3%	
fricatives	3.4%	18.3%	25.9%	21.1%	14.2%	7.3%	1.9%	0.9%
nombre d'oppositions de lieux	0	1	3	6	10	15	21	28

lieux d'articulation, il recrute les régions bilabiale, coronale, vélaire, uvulaire et glottale.
Une seule consonne nasale dans un système, elle est de type coronal. S'il en possède 2, elles sont de type bilabial et coronal. Réparties sur 3 lieux, s'ajoute une nasale vélaire. Puis à 4, apparaît le lieu palatal.

Si les fricatives d'un système n'exploitent qu'un lieu d'articulation, il s'agit de la région alvéolaire donc d'une coronale ; 2 lieux : alvéolaire et glottal ; à 3 lieux apparaissent les fricatives labiodentales ; à 4 est exploitée la région palatale et à 5 lieux apparaissent les fricatives vélaires.

Les affriquées, quant à elles, se répartissent dans les régions d'articulation des coronales, quel que soit la taille du système de lieu (1 à 3).

De la répartition des systèmes de lieux par catégories nous retenons les tendances suivantes :

- Comme pour les systèmes vocaliques [Val94], le système de lieux à n éléments contient le système à n-1 lieux consonantiques.
- Lorsqu'il n'y a qu'un lieu, quelle que soit la catégorie (plosives, fricatives ou autres), il s'agit de consonnes coronales (lieux 4 et 5).
- 95% des langues contrastent sur 3 lieux à l'intérieur des catégories (plus d'une langue sur 3 pour les nasales).
- Une taille de système domine chaque catégorie : 43% des langues ont 4 lieux pour les plosives et 74% ont 3 ou 4 lieux) ; 32% des langues répartissent leur nasales sur 3 lieux (59% sur 3 et 4) ; 28% présentent 3 lieux pour les fricatives (48% sur 3 et 4) ; 43% n'ont qu'un lieu d'articulation pour les affriquées.
- Les combinaisons à 5 lieux sont présentes dans seulement 36% des langues, essentiellement dans les plosives orales (97 langues).
- Les combinaisons à 6, 7 ou 8 lieux sont peu répandues (15% des langues). On les rencontre essentiellement dans les systèmes de fricatives.

Nos résultats confirment que les plosives sont de loin les phonèmes consonantiques « vedettes », omniprésentes dans les langues.

2.3.3 Rapport sourdes/sonores

Dans la catégorie des plosives orales, les sourdes dominent largement les sonores (64%) : on relève 30% de sourdes en plus dans les bilabiales, 50% de sourdes en plus dans les coronales, 60% dans la catégorie vélaire. Font exception à cette tendance, les langues du continent africain – familles niger-kordofanienne, nilo-saharienne et afro-asiatique : elles présentent en moyenne 1,5 fois plus de plosives bilabiales sonores que de sourdes.

Ce rapport sourdes/sonores est encore plus écrasant pour les fricatives : les langues présentent 5 fois plus de /s/ que de /z/, 3,2 fois plus de /ʃ/ que de /ʒ/ et presque 2 fois plus de /f/ que de /v/. Quel que soit le lieu d'articulation, le trait sourd concerne 72% des fricatives.

Les affriquées sont également majoritairement sourdes (74%).

3. UNIVERSAUX ET ONTOGENESE

Nous avons rassemblé les résultats de plusieurs études sur le contenu consonantique des productions à différents stades du babillage canonique (réduplication de syllabes identiques) ou panaché (avec variation des types de closions). L'ensemble de ces études couvre une période de 6-10 [Boy96] à 15-24 mois [Sto85] dans un environnement linguistique anglo-américain, français et de manière plus étendue sur 15 langues [Loc83]. [Boy96] relève que [p b m t d n k g ŋ] constituent 80% des productions « consonantiques » dans les premiers mois du babillage, [p b m t d] étant les plus fréquentes. La revue de question de [Loc83] livre par ordre décroissant de fréquence l'inventaire des consonnes représentant 95% des occurrences produites par des enfants âgés de 6 à 15 mois : [b] et [m], [p], [d], [h] et [n], [t], [g] et [k], [j] et [w], [s]. Les observations de [Rob94], sur une période plus longue (de 8 à 25 mois), montrent que chez les 8-12 mois plusieurs types de closions co-émergent. Les productions les plus fréquentes sont des plosives (orales et nasales), de lieu alvéolaire et bilabial. Ils confirment ainsi [Dav94], qui relèvent qu'au cours des 7-12 mois, ces consonnes couvrent 84 % des productions ([t d] 46%, [n] 18%, [p b] 15%, [m] 5%) ; les vélaires et glottales sont plus tardives. Enfin, l'étude de [Sto85] rend compte de l'apparition des segments consonantiques pour la période 15-24 mois : [b d m n j w h].

Table 3. Les catégories consonantiques (plosives, nasales, fricatives, affriquées) d'UPSID$_{451}$, classées par nombre de lieux d'articulation avec le pourcentage de langues correspondant (par exemple, 31% des 451 langues présentent un système de plosives réparties sur 3 lieux, 43% sur 4 lieux etc. ; 0 lieu correspondant à une absence de catégorie). Ne sont portés dans cette table que les systèmes les plus répandus.

lieux	plosives		nasales		fricatives		affriquées	
0	0%	–	3.5%	–	7%	–	33%	–
1	0%	–	1.8%	coronal	6.6%	coronal	42%	coronal
2	0.2%	coronal bilabial	31%	coronal bilabial	18%	coronal glottal	20%	coronal coronal 2
3	31%	coronal bilabial vélaire	32%	coronal bilabial vélaire	28%	coronal glottal labiodental	4.9%	coronal coronal 2 coronal 3
4	43%	coronal bilabial vélaire glottal	27%	coronal bilabial vélaire palatal	20%	coronal glottal labiodental coronal 2	–	–
5	21%	coronal bilabial vélaire glottal uvulaire	3.5%	coronal bilabial vélaire palatal coronal 2	11%	coronal glottal labiodental coronal 2 vélaire	–	–

Si on met en parallèle ces résultats sur l'acquisition du langage avec les données typologiques sur les langues du monde, on constate que les types de closions produits dans les premiers mois du babillage (vers 8-12 mois), ainsi que les sons acquis plus tardivement au cours de la deuxième année, correspondent aux phonèmes consonantiques les plus fréquents dans les langues (l'ordre d'émergence mis à part). On remarque aussi que les systèmes de lieux préférés, mis en évidence par mode dans l'analyse typologique, sont dominants dans les productions du babillage : alvéolaire et bilabial, devant vélaire et glottal, pour le mode plosif (oral ou nasal) ; alvéolaire et glottal ([s] et [h]) pour le mode fricatif qui est moins fréquent ; labiovélaire, coronal et palatal pour les approximantes [w], [l], [j], acquises plus tardivement.

Ces constatations nous permettent alors de formuler l'hypothèse que les sons des langues sont sans doute puisés dans le stock des capacités potentielles du babillage. [Loc92] n'hésitaient pas à conclure :

> « Infants heavily favor stop consonants over fricatives, and there are languages that have stops and no fricatives but no languages that exemplify the reverse pattern. [Such] "phonologically universal" patterns, which cut across languages and speakers are, in effect, the phonetic properties of *Homo sapiens*. » (p. 96)

L'*ontogenèse* contribue ainsi à fournir, à l'évidence, des éléments déterminants sur les dépendances entre certaines tendances générales des systèmes phonologiques et les capacités sensori-motrices de l'*Homo loquens*.

4. PERSPECTIVES

La typologie des systèmes consonantiques, mise ainsi en évidence, montre que les langues n'utilisent qu'un petit inventaire de consonnes (22 en moyenne) alors que potentiellement elles disposent d'un nombre considérable de possibilités (13 lieux x 7 catégories x les traits de mode). Les systèmes de lieux privilégiés pour chaque mode, montrent que le choix des langues est vraisemblablement conditionné, pour une bonne partie, par des contraintes morpho-physiologiques qui rendent faciles ou complexes, possibles ou non, les gestes selon les lieux (ex. la complexité du contrôle articulatoire des plosives voisées [Oha83] et l'aérodynamique des fricatives [Maw97].

Bien que plaidant la cause d'une linguistique non-substantialiste, Troubetzkoy concédait :

> « Ce fait ne peut être dû au hasard et doit avoir sa cause profonde dans la nature des trois séries en question. Le plus simple est de voir cette cause dans le fait que les lèvres, la pointe de la langue et le dos de la langue sont les organes mobiles les plus propres à obstruer l'espace buccal. » (1939, 135).

Mais en fait, c'est dans le cadre de la syllabe associant consonnes et voyelles, qu'il faut élaborer et mettre à l'épreuve des modèles de prédiction des tendances générales des structures sonores. Et en cela, nous rejoignons l'hypothèse de MacNeilage et Davis, selon laquelle l'émergence phonologique serait contrainte par le cadre syllabique [Mac90] [Mac98].

Remerciements : à Nourredine Abbadeni, Élisabeth Pinto et Florence Mistrulli, pour leur participation à l'implantation d'UPSID451. Cette recherche a été financée, en partie, par une action de la Région Rhône-Alpes (ARASSH) et du GIS des Sciences de la Cognition.

REFERENCES

[Boy96] Boysson-Bardies de B. (1996) *Comment la parole vient aux enfants*, Odile Jacob, Paris.

[Cre94] Creissels D. (1994) *Aperçus sur les structures phonologiques des langues négro-africaines*, 2ème éd., ELLUG, Grenoble, 322p.

[Cro78] Crothers J. (1978) Typology and universals of vowel systems in phonology, *Universals of Human Language*, Ed. by Greenberg H.J., Stanford University Press, Stanford, California, 2, 95-152.

[Dav94] Davis B.L., MacNeilage P.F. (1994) Oganization of babbling: A case study, *Language and Speech*, 37 (4), 341-355.

[Gre78] Greenberg J., Ferguson C., Moravcsik E.. (Ed.) (1978) *Universals of Human Languages: Method and Theory, Phonology, Word Structure, Syntax*, Stanford University Press.

[Hag82] Hagège C. (1982) *La structure des langues, collection* Que Sais-je ? PUF, Paris, 51-74

[Hoc55] Hockett C.F. (1955) *A Manual of Phonology*, Waverly Press, Baltimore.

[Kea90] Keating P.A. (1990) Coronal places of articulation, *UCLA WPP 74*, 35-60.

[Lad96] Ladefoged P., Maddieson I. (1996) The sounds of world languages, Blackwell Publishers, Oxford.

[Lav94] Lavers J. (1994) *Principles of Phonetics*, Cambridge University Press, Cambridge.

[Lin88] Lindblom B., Maddieson I. (1988) Phonetic universals in consonant systems, in *Language, Speech and Mind*, ed. by Hyman L.H. & Li C.N., London, New-york, 62-78.

[Loc83] Locke J.L. (1983) *Phonological acquisition and change*, Academic Press, New-York.

[Loc92] Locke J.L., Pearson D.M. (1992) Vocal learning and the emergence of phonological capacity. A neurobiological approach, in *Phonological development. Models, Research, Implications*, Ed. by C. Ferguson, L. Menn & C. Stoel-Gammon., Timomium, Maryland, 91-129.

[Mac90] MacNeilage P.F., Davis B.L. (1990) Acquisition of speech production: Achievement of segmental independence, in *Speech Production and Speech Modeling*, ed. by W.I. Hardcastle & A. Marchal, Kluwer, Dordrecht.

[Mac98] MacNeilage P.F. (1998) The Frame/Content theory of evolution of speech production, *Behavioral and Brain Sciences*, à paraître.

[Mad86] Maddieson I. (1986) *Patterns of sounds*, Cambridge University Press, Cambrige.

[Mad89] Maddieson I., Precoda K. (1989) Updating UPSID, *UCLA WPP* 74, 104-111.

[Maw97] Mawass K. (1997) *Synthèse articulatoire des consonnes fricatives du français*, Thèse de doctorat, INPG, Grenoble.

[Oha83] Ohala J. (1983) The origin of sound patterns in vocal tract constraints, *The Production of Speech*, ed. by P.F. MacNeilage, Springer Verlag, New-York, 189-216.

[Rob94] Robb M., Bleile K. (1994) Consonant inventories of young children from 8 to 25 months, *Clinical Linguistics and Phonetics*, 8, 295-320.

[Tro39] Troubetzkoy N.S. (1939). *Grundzüge des Phonologie*, Travaux du Cercle Linguistique de Prague 7, 272p. Traduit en français par Cantineau J. (1970) sous le titre : Principes de phonologie, Klincksieck, Paris, 394p.

[Sed69] Sedlak P. (1969) Typological considerations of vowel quality systems, *Working Papers on Language Universals*, 1, Stanford University, 1-40.

[Ste96] Stefanuto M. (1996) *Typologie des lieux d'articulation des langues du monde*, TER Sciences du Langage, Université Stendhal, Grenoble.

[Sto85] Stoel-Gammon C. (1985) Phonetic Inventories 15-24 months: A longitudinal study, *J. of Speech and Hearing Research*, 23, 506-512.

[Val94] Vallée N. (1994) *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de Doctorat en Sciences du Langage, Université Stendhal, Grenoble.

[i a u] ? À la portée d'un conduit vocal de Néandertal

Louis-Jean BOË[1] Shinji MAEDA[2] Christian ABRY[1] Jean-Louis HEIM[3]

1. ICP
INPG/Université Stendhal
BP 25 38040 Grenoble Cedex
abry,boe@icp.inpg.fr

2. ENST
46, rue Barrault
75 013 Paris
maeda@enst.fr

3. Laboratoire d'Anthropologie
17, place du Trocadero
75 116 Paris
heim@mnhm.fr

ABSTRACT

Since Lieberman & Crelin (1971) it has been repeatedly quoted by many scholars in different scientific areas, in connection with linguistics, that it was impossible for Neandertal men to produce the cardinal vowels [i a u], encountered in nearly all languages in the world .

Using an articulatory model, the main flaws of Lieberman & Crelin's presuppositions, reconstructions and computations, will be demonstrated. Our demonstration will by no means prove whether Neandertal men could speak or not, but simply that their peripheral abilities, i.e. vocal tract per se, could not have prevented them from contrasting such universal vowels.

1. UNE BASE ANATOMIQUE NÉCESSAIRE POUR LA PAROLE ?

ONTOGENÈSE ET PHYLOGENÈSE

Une question très controversée et alimentée depuis plus de 30 ans par des anthropologues, des paléontologues, des anatomistes, des linguistes et des spécialistes de la parole est la suivante : les Néandertaliens pouvaient-ils produire des sons tels que ceux que l'on répertorie maintenant dans les langues du monde ? Plus précisément pouvaient-ils articuler [i a u], les voyelles à contraste maximal quasiment omniprésent ?

En dépit des réserves émises par les plus grands spécialistes (de Leroi-Gourhan à Pinker) on trouve maintes fois citée et systématiquement diffusée dans des publications, dans des encyclopédies, dans des revues de vulgarisation (*Scientific American, Pour la Science, La Recherche*), qui traitent ou font référence à l'émergence de la parole, la thèse de Lieberman & Crelin présentée à de nombreuses reprises depuis 1971 [Lie71-91] et soutenue par Laitman [Lai83-90]. Selon celle-ci, pour que la production de la parole soit devenue possible au cours de l'évolution, il a fallu que la base du crâne fléchisse suffisamment, que le larynx descende (pour un homme, en regard de la quatrième vertèbre cervicale C4), ceci libérant une cavité pharyngale suffisamment volumineuse. C'est à cette condition que l'espace acoustique des sons serait potentiellement assez étendu pour réaliser les contrastes vocaliques de type [i a u]. L'acquisition de cette *base anatomique* aurait été la condition nécessaire – associée à des développements corticaux et cognitifs importants – pour que nos ancêtres puissent articuler ces sons de parole.

Selon cette thèse, les Néandertaliens ne possédaient pas cette base anatomique, ils ne pouvaient donc pas parler, ce qui a vraisemblablement entraîné leur disparition. Lieberman & Crelin, notant les similarités anatomiques du conduit vocal entre le nouveau-né, le chimpanzé et le Néandertalien, ont adopté la thèse de la récapitulation : l'ontogenèse reflète la phylogenèse, (la loi de Haeckel de 1866). Le nouveau-né humain présente un larynx en position haute, ce qui diminue les risques d'étouffement , il peut ainsi avaler tout en respirant ; en contrepartie il ne peut, pas davantage que le singe, produire des voyelles. Puis son larynx s'abaisse progressivement et, à partir d'un an, son pharynx est suffisamment volumineux pour qu'il puisse commencer à contraster [i a u]. Cette croissance du conduit vocal, favorisant le pharynx plus que la cavité buccale, reflète bien, d'après Lieberman & Crelin, la phylogenèse d'*Homo sapiens*.

Sur la base anatomique de la position du larynx Liberman & Crelin opposeront d'une part le singe, le Néandertalien et le nouveau-né, d'autre part l'Homme moderne qui est le seul, des espèces vivantes, à parler.

2. LA RECONSTRUCTION DU CONDUIT VOCAL DES NÉANDERTALIENS

Nous retiendrons de l'anatomie du crâne et de la mâchoire des Néandertaliens les informations qui permettent d'inférer, dans une certaine mesure, la position du larynx. Au début du XX^{e} siècle, ceux-ci étaient considérés comme beaucoup plus proche des singes que de l'Homme [Hei86]. Boule, la plus grande autorité française de cette époque, ne disposait que de très peu d'éléments de comparaison pour reconstituer avec précision la base du crâne de l'*Homme de la Chapelle-aux-Saints* [Bou11]. Pour la rendre compatible avec la position du corps qu'il supposait fléchie, il a déterminé le centre de gravité de la tête trop antérieurement en donnant au basion et au trou occipital une position exagérément reculée (environ 4mm). Malgré son insistance à mettre l'accent sur les caractères primitifs des Néandertaliens, Boule a bien noté que les dispositions des apophyses géni et de celles correspondant au digastrique n'étaient pas très différentes de celles de l'Homme moderne [Bou23].

Dès 1984, Heim a proposé une nouvelle reconstitution du crâne de l'*Homme de la Chapelle-aux-Saints* en tenant compte de rapports anatomiques réels. Il a mis en évidence que la position de la tête et la base du crâne ne différaient guère fondamentalement de celle d'un homme d'aujourd'hui par leur attitude au-dessus de la colonne vertébrale. L'angle de Landzert de sa reconstitution est de 137.5° (au lieu de 150° pour celle de Boule), ce qui correspond bien aux valeurs établies pour le Néandertalien de *Gibraltar* (142°), celui de *La Ferrassie 1* (135°) et de l'Homme (entre 126° et 143°). Le larynx du Néandertalien devait, selon toute vraisemblance, occuper la même position que celle des Hommes modernes [Hei89, 90].

C'est exactement à la même conclusion que sont arrivés, la même année, les chercheurs d'une équipe pluridisciplinaire (université de Bordeaux I, CNRS, université de Tel Aviv, Moorehead State University aux USA), après la découverte, publiée en 1985, d'un squelette de Néandertalien vieux de 60.000 ans (phase tardive du moustérien), reposant dans une sépulture du site de Kebara en Israël. Pour la première fois un os hyoïde était retrouvé, préservé dans son intégralité. Le fait qu'il soit identique en taille et en forme à celui des

hommes d'aujourd'hui, ainsi que l'analyse de la disposition des insertions musculaires de la mandibule les ont conduit à la conclusion que « there has been little or no change in the visceral skeleton (including the hyoid [...] and inferentially the larynx) during the past 60,000 years of human evolution » [Are89].

En 1971 Liberman & Crelin, pour reconstruire le conduit vocal du Néandertalien de *la Chapelle-aux-Saints*, ne pouvaient certes qu'utiliser le moulage du crâne selon Boule. La position du larynx, telle qu'ils l'ont inférée, était ainsi nettement trop haute. Toute leur démonstration sur la classification des Néandertaliens en fonction de cette base anatomique en est donc aussi faussée.

3. LA SIMULATION ACOUSTIQUE DE LIBERMAN & CRELIN

Rappelons brièvement les points clés de leur démonstration. Après une reconstruction discutable du conduit vocal néandertalien à partir du crâne de l'*Homme de la Chapelle-aux-Saints* et la production d'un moulage, Liberman & Crelin en ont estimé la fonction d'aire. Puis il l'ont modifiée en 3 ou 4 étapes pour produire les trois voyelles [i a u] [Lie72, fig. 28].

Toute la démonstration ne repose que sur trois fonctions d'aire proposées comme prototypes de ce que les Néandertaliens étaient supposés faire de mieux comme contraste vocalique maximal.

Une analyse de ces fonctions d'aire montre, à l'évidence, que Lieberman & Crelin ne pouvaient obtenir que des *voyelles centrales* et non de véritables cardinales extrêmes : les constrictions ne sont pas suffisantes, *mais surtout elles ne sont pas assez longues*. La démonstration repose donc sur une détermination biaisée de trois fonctions d'aire.

4. MODÉLISATION DE LA CROISSANCE ET PROTOTYPES VOCALIQUES

4.1 Modélisation de la croissance

Pour les locuteurs adultes on dispose, depuis une vingtaine d'années, de nombreux modèles articulatoires du conduit vocal établis à partir de données cinéradiographiques et dérivés d'une analyse statistique guidée par des connaissances de la physiologie des articulateurs, de Maeda [Mae79, 90] à Vilain [Vil97). Ces modèles présentent l'avantage de rendre compte intrinsèquement de certaines contraintes articulatoires de production : les paramètres de commande sont directement interprétables en termes de degrés de liberté des articulateurs. Mais rares sont ceux qui possèdent de telles caractéristiques tout en simulant la croissance, de la naissance à l'âge adulte.

Nous avons utilisé le modèle de croissance VLAM (*Variable Linear Articulatory Model*) – développé par Maeda en 1994 [Boë97] – qui intègre les connaissances acquises dans des modèles précédents et les données de croissance actuellement disponibles. Il est en effet possible de tirer parti des mesures crânio-faciales établies à des âges différents et publiées en anatomie, radiologie et pédiatrie. L'évolution des dimensions de la tête (l'architecture osseuse) et celle de la position de l'os hyoïde (dans une certaine mesure associée à la position du larynx) permettent d'inférer les grandes tendances du développement du conduit vocal. La thèse de Goldstein présente, sur ce point, une véritable mine d'informations, avec un inventaire de données correspondant à 14 distances et 3 mesures angulaires établies par rapport à des repères et des lignes de référence anatomiques, pour des âges allant de quelques mois à 20 ans [Gol80]. Toutes ces données ont été ajustées par des (doubles) sigmoïdes qui caractérisent bien la croissance squelettale et musculaire en général. À la naissance, l'enfant présente une tête plutôt hémisphérique. Avec l'âge, les augmentations du volume et de la forme du crâne, de la taille du maxillaire inférieur vont modifier les proportions entre les dimensions verticale et horizontale. Il ne s'agit pas d'une homothétie pure et simple mais d'une anamorphose dans laquelle la croissance du pharynx est pratiquement deux fois plus importante que celle de la cavité avant.

Dans VLAM, l'évolution de la croissance a été introduite par Maeda dans la dimension longitudinale du conduit vocal par deux facteurs d'échelle, l'un pour la partie antérieure du conduit vocal et l'autre pour le pharynx, la zone intermédiaire étant interpolée :

$$\text{pharynx_scale} = k(1.0 - 0.30) + 0.30$$
$$\text{mouth_scale} = k(1.0 - 0.65) + 0.65$$

Le facteur k permet ainsi de simuler l'évolution du conduit vocal, mois par mois, année par année : il a été étalonné en âge grâce aux données de Goldstein [Boë97]. Avec VLAM il est donc possible de simuler une croissance linéaire qui conserverait au conduit vocal les mêmes proportions, de la naissance à l'âge adulte, comme dans le cas du singe, ou du Néandertalien selon [Lie71]. Le modèle VLAM a été implanté à l'ICP dans un l'environnement *Growth* adapté à des simulations systématiques.

4.2 Espace maximal et prototypes

Les modèles délivrent une coupe sagittale en 2D et la fonction d'aire correspondante à partir de laquelle il est possible de calculer la réponse harmonique, les formants et le son. Cette procédure est bien adaptée à la modélisation des productions vocaliques. Si l'on explore, en entrée, tout l'espace possible des paramètres de commande – tout en restant dans des conditions de production vocalique – on peut simuler, en sortie, l'espace acoustique maximal F_1 F_2 F_3. Toutes les voyelles orales possibles sont ainsi localisables dans ces limites [Boë89]. Ce type de génération en extension permet de décrire avec précision les potentialités de distinctivité maximale et de procéder à un choix optimal de réalisations prototypiques. Une telle démarche se révèle plus sûre que celle qui consiste à extrapoler les limites de l'espace vocalique d'un conduit vocal (ou d'un modèle) à partir de trois uniques exemplaires [i a u] dont il n'est pas certain qu'ils soient optimaux, comme nous l'avons vu en discutant les résultats de [Lie71].

Avec VLAM, nous avons généré un ensemble de voyelles pour un maillage des paramètres de commande ($-2 \leq P_i \leq +2$ avec un pas de 0.5), avec des contraintes sur les constrictions intrabuccale et labiale conservées identiques pour le nouveau-né et l'adulte (aire de la constriction ≥ 0.3 cm^2 et aire aux lèvres ≥ 0.15 cm^2). Le Tableau 1 présente les limites de cet espace (en Hz) et dans des échelles perceptives (en Bark et en ERB rate). On constate que l'espace acoustico-perceptif F_1 et F_2 est au moins aussi important, si ce n'est plus, pour les nouveau-nés que pour les adultes, en revanche celui de F_3 est peu réduit (de 3.2 Bark ou 4.3 ERB rate).

Grâce à cet espace maximal nous avons établi les prototypes pour les voyelles [i a u] en tenant compte des propositions de Goldstein. Le modèle VLAM nous

a permis ainsi de calculer : ① les configurations articulatoires qu'aurait un nouveau-né s'il manifestait les mêmes capacités de contrôle qu'un adulte ; ② les configurations pour un adulte ; ③ celles d'un adulte qui aurait eu une croissance du conduit vocal linéaire, sans normalisation de longueur ; ④ avec normalisation de longueur (Tableau 2). On peut vérifier les tendances déjà bien notées au Tableau 1, à savoir dans tous les cas un espace acoustico-perceptif F_1 et F_2 quasi identique et un espace F_3, plus important pour le conduit vocal de l'adulte humain.

Le Tableau 3 présente les contrastes maximaux entre [i], [a] et [u] soit F_1[a] - F_1[i], F_2[i] - F_2[u], F_3[i] - F_3[u]. On constate bien :

- les mêmes contrastes perceptifs entre l'enfant et l'adulte pour F_1 et F_2 : F_1[i] et F_1[a] dépendent de la cavité d'avant ; par le jeu des affiliations, F_2[i] et F_2[u] se retrouvent aussi associés, chez l'enfant, à la cavité d'avant ;
- un contraste F_2 F_3 pour [i] moins important chez l'enfant que chez l'adulte, ces deux formants devenant associés à des cavités avant et arrière de longueur assez voisines (entre 3 et 4 cm).

Les relations établies entre fonctions d'aires, cavités et formants (Tableau 4) permettent de bien expliquer ces tendances [Boë98]. Le patron formantique de [i] nécessite une composante fréquentielle élevée correspondant à la longueur de la constriction d'avant, soit F3 ; il requiert aussi que la longueur de la cavité arrière soit suffisamment courte pour faire converger F_2 et F_3 dans cette même gamme de fréquences. Pour une constriction d'avant donnée, le volume de la cavité d'arrière n'est donc responsable que de la fréquence basse de F_1. Mais, même si le F_1[i] d'un enfant était égal au F1[a] d'un adulte (environ 800 Hz), il y aurait quand même un contraste voyelle ouverte / voyelle fermée d'environ 5 Bark (ou 6.5 ERB), puisque le F_1[a] du bébé est toujours élevé (entre 1400 et 1600 Hz), quelle que soit son affiliation à la cavité avant ou à la cavité arrière. Ce n'est donc clairement pas à cause du faible volume de leur pharynx que les nouveau-nés ne produisent pas [i] et [u] : il faut, pour en trouver la cause, se tourner du côté du contrôle moteur [Mac98, Abr98], en l'occurrence vraisemblablement vers la difficulté à contrôler les constrictions de ces voyelles sans produire d'occlusion.

5. CONCLUSION

Un conduit vocal présentant un larynx en position haute et donc un petit pharynx, comme celui du nouveau-né, peut produire un espace vocalique tout aussi important qu'un adulte ayant un larynx en position basse et donc un pharynx plus volumineux.

Si un homme devenu adulte avait son larynx en position haute à la suite d'une croissance linéaire de son conduit vocal, il aurait de la même manière, un espace vocalique tout aussi important.

En conséquence, même si les Néandertaliens avaient un larynx en position haute, ce qui n'est vraisemblablement le cas selon [Are89] et [Hei89], ils auraient pu produire des voyelles comme l'Homme moderne.

Ou : comment une démonstration reposant sur une reconstitution articulatoire biaisée et une simulation acoustique erronée dès le départ a pu faire si longtemps illusion ? Mais ceci est une autre histoire... qui relève de la sociologie des sciences.

RÉFÉRENCES

[Abr98] Abry C., Boë L.-J., Laboissière R., Schwartz J.-L. (1998) Emergence of a new puzzle for the evolution of speech and language? *Commentaire à* [Mac98], à paraître.

[Are89] Arensburg B. et al. (1989) A middle paleolithic human hyoid bone. *Nature* 338, 758-760.

[Boë97] Boë L.-J., Maeda S. (1997) Modélisation de la croissance du conduit vocal. Espaces vocaliques des nouveaux-nés et des adultes. Conséquences pour la phylogenèse et l'ontogenèse. *La voyelle dans tous ses états*, 98-105.

[Boë89] Boë L.J., Perrier P., Guérin B., Schwartz J.L. (1989) Maximal Vowel Space. *Eurospeech* 89, 2, 281-284.

[Bou11] Boule M. (1911-1913) L'homme fossile de la Chapelle-aux-Saints. *Annales de Paléantologie* 6, 109 ; 7, 21-85 ; 8, 1.

[Bou23] Boule M. (1923) *Les hommes fossiles. Éléments de paléontologie humaine.* Paris, Masson. 2e édition.

[Gol80] Goldstein U.G. (1980) *An articulatory model for the vocal tract of the growing children.* Thesis of Doctor of Science, MIT, Cambridge, Massachusett.

[Hei86] Heim J.-L. (1989) *La reconstitution du crâne de la Chapelle-aux-Saints.* Film 16mm, Paris, Éd. SFRS.

[Hei88] Heim J.-L. (1988) *L'évolution dans sa réalité.* Paris, Masson, 179-211.

[Hei89] Heim J.-L. (1989) Une nouvelle reconstitution du crâne néandertalien de la Chapelle-aux-Saints. *C.R. Acad. Sc. Paris*, t. 308, série II, n° 6, 1187-1192.

[Hei90] Heim J.-L. (1990) La nouvelle reconstitution du crâne néandertalien de la Chapelle-aux-Saints. Méthode et résultats. *Bull. et Mém. de la Soc. d'Anthropol. de Paris* 6, 1-2, 94-117.

[Lai83] Laitman J.T. (1983) *The evolution of hominid upper respiratory system and implication for the origin of* speech. Glossogenetics 63-90. Harwood Academic publishers.

[Lai90] Laitman J.T. et al. (1990) The Kebara hyoid : What can it tell us about the evolution of the hominid vocal-tract ? *Am. J. of Physical Anthropolgy* 81, 254.

[Lie71] Lieberman Ph., Crelin E.S. (1971) On the speech of the Neanderthal man. *Linguistic Inquiry* 2, 2, 203-222.

[Lie72] Lieberman Ph. (1972) *The speech of primates.* La Haye, Mouton.

[Lie91] Lieberman Ph. (1991) *Uniquely human. The evolution of speech, thought, and selfless behaviour.* Cambridge, Mass., Harvard University Press.

[Mae79] Maeda S. (1979) Un modèle articulatoire basé sur une étude acoustique. *Bull. de l'Institut de Phonétique de Grenoble* VIII, 35-55.

[Mae89] Maeda S. (1989) *Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model.* In Speech Production and Modelling, 131-149. Acad. Pub., Kluwer.

[Mac98] MacNeilage P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, à paraître.

[Vil97] Vilain A. (1997) *Un nouveau modèle articulatoire pour la synthèse et le contrôle robotique de la parole : Gentiane.* DEA Sci. du Langage, Univ. Stendhal, Grenoble.

Remerciements : Cette recherche a été financée, en partie, par une action de la Région Rhône-Alpes (ARASSH) et du GIS des Sciences de la Cognition.

Tableau 1. Limites F_1 F_2 F_3 de l'espace maximal établies avec le modèle VLAM en Hz et en unités perceptives : Bark et ERB rate.

	F_1 max – F_1 min (Hz	Bark	ERB)	F_2 max – F_2 min (Hz	Bark	ERB)	F_3 max – F_3 min (Hz	Bark	ERB)
adulte	546	4.7	7.5	1387	6.2	8.1	2018	5.2	6.4
nouveau-né	1302	7.1	9.5	3246	7.2	8.7	1658	2.0	2.1

Tableau 2. Prototypes [i a u] établis avec le modèle VLAM. Longueur du conduit vocal (en cm), formants (en Hz, en Bark et ERB rate).

	L (cm)	F_1 (Hz	Bark	ERB)	F2 (Hz	Bark	ERB)	F3 (Hz	Bark	ERB)
[i]										
① nouveau-né	7.7	628	6.0	11.8	4824	18.9	28.1	5714	20.1	29.3
② adulte	17.4	246	2.6	6.3	2058	13.1	21.2	3325	16.3	25.1
③ homothétie linéaire	13.7	360	3.7	8.3	2902	15.4	24.0	3474	16.6	25.5
④ homothétie + norm	17.4	306	3.2	7.4	2293	13.8	22.1	2846	15.3	23.9
[a]										
① nouveau-né	7.3	1489	11.0	18.5	3477	16.7	25.5	5430	19.7	29.0
② adulte	17.3	678	6.4	12.4	1275	10.0	17.2	2217	13.6	22.8
③ homothétie linéaire	13.0	913	8.0	14.6	1817	12.3	20.1	3041	15.7	24.4
④ homothétie + norm	17.3	720	6.7	12.8	1383	10.5	18.0	2304	13.8	22.1
[u]										
① nouveau-né	8.6	538	5.3	10.8	1614	11.5	19.1	5356	19.6	28.9
② adulte	19.4	299	3.1	7.3	682	6.4	12.4	1985	12.8	20.9
③ homothétie linéaire	15.1	365	3.7	8.4	889	7.8	14.4	2913	15.4	24.0
④ homothétie + norm	19.4	312	3.2	7.5	706	6.6	12.7	2267	13.7	22.0

Tableau 3. Contrastes entre [i] [a] et [u] , dans les dimensions F_1 F_2 F_3 (en Bark et en ERB rate) correspondant aux prototypes établis avec le modèle de croissance VLAM.

	F_1 [a] – F_1 [i] (Bark	ERB rate)	F_2 [i] – F_2 [u] (Bark	ERB rate)	F_3 [i] – F_3 [u] (Bark	ERB rate)
① nouveau-né	5.0	6.7	7.4	9.0	0.5	0.4
② adulte	3.8	6.1	6.7	8.8	3.5	4.2
③ homothétie linéaire	4.3	6.3	7.6	9.6	1.2	1.5
④ homothétie + norm	3.5	5.4	7.2	9.4	1.6	1.9

Tableau 4. Affiliations des cavités et des formants pour des prototypes du français et pour un conduit vocal d'Homme adulte.

	F_1	F_2	F_3
[i]	Helmholtz : pharynx + constriction	λ/2 pharynx	λ/2 constriction avant
[a]	λ/4 cavité buccale	λ/4 pharynx	3λ/4 cavité buccale
[u]	Helmholtz avant : lèvres + cavité buccale	Helmholtz arrière : constriction + cavité arrière	λ/2 cavité arrière

Conférence invitée

L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole

K. R. Scherer, T. Johnstone, J. Sangsue

FPSE, Université de Genève
9, Route de Drize, 1227 Carouge, Suisse
e-mail: scherer@uni2a.unige.ch

"Il y a autant de mouvements dans la voix qu'il y a de mouvements dans l'esprit, et l'esprit est profondément affecté par la voix."

Cicéron, De Oratore

Abstract

This contribution argues that speech technologies, specifically speaker verification, speech recognition and speech synthesis, need to model the effects of speaker state (attitudes and emotions) in order to increase their quality and acceptance. Experimental research on the effects of stress and emotion on voice quality and prosody is reviewed and linked to basic dimensions of speech communication. It is concluded that the current state of the art in this area can provide speech technologists with important leads for modeling affective speaker states. We argue that there is a strong need for increased collaboration between speech engineers and speech scientists from other disciplines.

1. Introduction

Les développements récents dans le domaine du traitement des signaux ont permis l'essor d'un grand nombre de technologies de la parole dont beaucoup sont actuellement utilisées, voire sur le point d'être commercialisées. Cependant, leur fiabilité et leur usage sont fortement limités par le fait que la parole est affectée par des changements transitoires de l'état du locuteur, changements liés à son état émotionnel, à ses attitudes, à sa santé ou pouvant se produire sous l'impact de stress cognitif ou émotionnel.

Quoi qu'il soit compréhensible que les chercheurs et les ingénieurs en technologie du langage aient donné la priorité au développement d'algorithmes de base pour l'analyse et la synthèse de la parole, nous soutiendrons que continuer à négliger les facteurs affectifs mènera l'utilisateur à refuser cette technologie et privera cette dernière d'une entrée dans le marché à laquelle elle pourrait potentiellement prétendre.

La synthèse de la parole fournit un exemple parlant des effets d'une négligence des aspects pragmatiques de la parole, en particulier l'influence des attitudes et des émotions. En effet, bien qu'une technologie de synthèse raisonnablement accessible existe depuis un certain nombre d'années, ce qu'il en ressort actuellement est fréquemment jugé comme inacceptable pour la plupart des applications dans le monde réel. La raison est que la majorité des systèmes TTS (*text-to-speech*), entre autres, sont limités dans la manière dont ils prennent en compte les caractéristiques prosodiques de base pour différents types de phrases, comme les interrogations. Ils échouent même à fournir d'importants indicateurs stylistiques et affectifs tels que la politesse, l'intérêt, le plaisir ou, à d'autres fins, la détermination ou l'insistance. Nous croyons qu'il est essentiel que les ingénieurs travaillant sur la synthèse deviennent plus conscients de l'importance des indices pragmatiques et émotionnels et tentent de construire des outils appropriés pour communiquer cette signification pragmatique et affective.

Les utilisateurs du monde réel demanderont d'ailleurs eux-mêmes que les techniques de synthèse de la parole fournissent de tels indices avant d'accepter cette technologie à grande échelle. Evidemment, le développement d'instruments adéquats et de systèmes de règles exigera un regard scientifique sur les propriétés prosodiques et paralinguistiques des actes de la parole, des émotions et des attitudes du locuteur. Malheureusement, ce terrain de recherche a été plutôt négligé dans le passé par les ingénieurs de la parole, les linguistes et les phonéticiens, tout comme par les psychologues.

La vérification automatique du locuteur est un autre exemple de technologie de la parole qui pourrait être utilement appliquée dans de nombreux secteurs économiques. L'automatisation progressivement croissante de beaucoup d'activités dans le domaine du service, spécialement l'accès télécommandé à l'information, le paiement automatique par carte de crédit et la consultation bancaire par téléphone (*telephone banking*), requièrent de plus en plus de moyens fiables pour assurer toute sécurité et limiter l'accès aux utilisateurs autorisés uniquement. Malgré des investissements conséquents et des tests en laboratoire prometteurs, la performance et l'acceptation de tels systèmes restent bien en-deça des attentes. Les systèmes de vérification du locuteur continuent à souffrir d'un nombre inacceptable d'erreurs, sous forme de faux rejets ou fausses identifications, et n'apparaissent pas suffisamment robustes pour un usage flexible dans les situations quotidiennes. En particulier, la performance de ces systèmes se dégrade avec le laps de temps entre l'entraînement et l'usage, un phénomène qui est probablement du, en majeure partie, à un changement d'état du locuteur [Dod98; Kar98].

Nous croyons que ces problèmes de performance et d'acceptation mentionnés ci-dessus sont partiellement dus au fait que beaucoup d'ingénieurs de la parole ou de scientifiques qui ont été particulièrement impliqués dans le développement de cette technologie ont négligé ou tout

au moins ont sous-estimé les effets puissants des variations transitoires de l'état organique du locuteur sur ses productions vocales et, en conséquence, sur les paramètres acoustiques utilisés dans les algorithmes de ces technologies du langage. Il est une croyance largement répandue dans la communauté scientifique: le fait d'augmenter davantage la sophistication des algorithmes statistiques permettant la reconnaissance et la comparaison des patterns résoudrait les problèmes posés par les variations internes au locuteur, variations produites par des changements attitudinaux ou émotionnels de son état et qui jusqu'à présent ont été traitées essentiellement au hasard.

Cependant, à mesure que le temps passe, il devient de plus en plus difficile de nier la nature asymptotique des progrès dans ce domaine et il est actuellement de moins en moins probable qu'une augmentation de la sophistication par une approche purement statistique mènera à des résultats persuasifs. L'importance des déficits de performance des systèmes de vérification du locuteur concernant la variation intra-locuteur n'a toutefois jamais encore été établie puisqu'il n'existe actuellement, à notre connaissance, aucune base de données avec des échantillons contenant des variations intra-locuteur systématiquement récoltées, qui pourraient être utilisés pour comparer les différents systèmes.

Les recherches antérieures ont largement démontré que les changements d'états du locuteur modifient les paramètres acoustiques de sa parole de façon spécifique et consistante [pour une revue de la littérature dans ce domaine, voir par exemple Fri77, Mur88, Pitt93; Sch81; Sch86; Sch95; Sch90]. Incorporer à la technologie de vérification du locuteur cette connaissance des dépendances à un état organique pourrait alors mener à des systèmes plus robustes, en permettant à la fois la sélection d'ensemble de caractéristiques acoustiques plus appropriées (traitant les algorithmes et éprouvant les procédures) et le développement de techniques de production de la parole élaborée de façon à minimiser les effets des changements transitoires de l'état physiologique du locuteur.

Une certaine partie de la communauté scientifique travaillant sur la parole, principalement les chercheurs orientés vers une approche multidisciplinaire de la technologie du langage, a insisté pendant longtemps sur le fait qu'une percée réelle dans la performance et l'acceptation des technologies de la parole ne peut être réalisée qu'en basant les développements ultérieurs sur une connaissance scientifique plus approfondie de la régularité et des lois des variations intra- et inter-locuteurs (voir, par exemple, les commentaires conclusifs de Fur94) ainsi que sur le rôle des indices pragmatiques dans la communication verbale. Un développement de notre compréhension des processus impliqués nécessite des travaux expérimentaux plus importants introduisant une manipulation systématique des états cognitifs, affectifs et attitudinaux du locuteur dans des populations distinctes de locuteurs. Ceci n'a jamais encore été systématiquement fait.

Notre groupe de recherche participe quelque temps aux efforts déployés pour étudier les effets des changements engendrés par différents types d'états physiologiques du locuteur sur les paramètres acoustiques de sa voix. Nous avons étudié les influences à court et long terme du stress, de l'émotion, des attitudes et des mensonges sur la voix et la parole, en utilisant des méthodes de traitement digital des signaux pour l'extraction de paramètres [Ban96; Lad85; Sch77; Sch84; Sch91; Tol86; Wal86]. Ceci a supposé le développement de techniques d'induction fiable et réaliste d'affects et de stress, basées sur des modèles théoriques testés empiriquement concernant les changements dans la voix suite à un changement affectif. Actuellement, nous poursuivons nos travaux en examinant les effets vocaux et acoustiques de l'induction d'événements émotionnels dans le cadre de jeu vidéo et autres interactions homme-ordinateur [Joh96]. Dans ce qui suit, nous soulignerons certaines questions et certains résultats que nous croyons pertinents pour une avancée dans la sophistication des technologies du langage.

Nous commencerons par exposer les principes de base sur la nature de la communication vocale qu'il s'agit de garder à l'esprit quoique travaillant sur des questions appliquées, plus moléculaires. Il est utile de débuter avec la communication animale, étant donné la continuité phylogénétique dans l'expression vocale des affects [Sch88].

2. Fonctions et déterminants multiples des vocalisations

Si l'expression vocale chez les animaux a d'abord été considérée comme un indicateur de leur état affectif ou motivationnel sous-jacent, les recherches plus récentes indiquent que la situation est plus complexe. Marler et ses collègues [Mar84], en étudiant les appels d'alarme chez une espèce de singe, ont trouvé que les appels ne sont pas seulement des indicateurs de leur peur mais qu'ils sont également liés spécifiquement aux types de prédateurs. Les cris d'alarme produits pour les léopards, les aigles ou les serpents, par exemple, présentent des sons, niveaux d'énergie et fréquences différents. C'est pourquoi, les auteurs ont rejeté l'hypothèse selon laquelle la communication animale serait limitée à n'indiquer que l'état émotionnel ou motivationnel de l'animal, en postulant que la plupart des appels des animaux ont un composant symbolique très fortement référenciel, partiellement appris. Le système d'alarme n'externaliserait pas seulement l'affect sous-jacent mais refléterait encore la classification du prédateur sur la base de processus cognitifs rudimentaires.

Mettre un accent exclusif soit sur l'expression de l'affect ou de la motivation, soit sur sa fonction symbolique, c'est omettre le fait que la plupart des signaux vocaux sont en fait pluri-fonctionnels. Le modèle " Organon " développé par Bühler [Büh34] peut être évoqué pour analyser les fonctions distinctes des affects vocaux. Dans ce modèle, un signe possède trois fonctions: il est un symbole, en représentant l'objet, l'événement ou le fait; il est un symptôme de l'état de l'utilisateur; il est enfin un appel ou un signal en ce qu'il tente de susciter une réponse du

récepteur. L'appel d'alarme du singe notamment comprend ces trois aspects : en tant que symbole de différents prédateurs, en tant que symptôme de l'état de peur de l'animal et en tant qu'appel aux autres membres de la troupe pour qu'ils s'enfuient. Par ailleurs, ces fonctions sont mutuellement inter-dépendantes. Si un appel se rapporte à un prédateur des airs, la réaction et le signal seront différents de ceux qui seraient faits en référence à un prédateur terrestre: dans le premier cas, l'émetteur et le récepteur tous deux devraient chercher une cachette sous des branches ou des buissons; dans le second cas, ils devraient s'activer physiologiquement et grimper à un arbre.

Outre ces multiples fonctions, nous devons considérer de multiples déterminants. Nous avons suggéré de distinguer les termes " effet push " et " effet pull ". Les effets " push " concernent les processus physiologiques, telle que la tension musculaire, qui " poussent " les vocalisations dans une certaine direction. Ils reflètent les changements dans les sous-systèmes physiologiques de l'organisme et ayant un effet direct sur les paramètres vocaux. Les effets " pull " ont trait, quant à eux, à des facteurs externes, telles les attentes de l'auditeur, qui " tirent " la vocalisation de l'affect vers un modèle acoustique particulier. Les facteurs " pull ", bien que médiatisés par des systèmes internes, sont basés sur l'extérieur. Ils opèrent vers une production de patterns ou de modèles acoustiques spécifiques, définis et valorisés socialement. Cette distinction est importante pour la compréhension des différences entre les productions vocales.

Ainsi, l'information dans la voix reflète aussi bien les modifications physiologiques, liées à l'état organique, et totalement involontaires qui affectent les systèmes de production du langage du locuteur (effet " push ") que l'adoption de styles de langage culturellement acceptés (effet " pull ") [Sch86]. Cette notion de changements involontaires et volontaires dans la production de la parole est cruciale pour une vérification du locuteur. Les modifications physiologiques à la base de la glotte sont susceptibles de produire des changements plutôt à long terme, supra-segmentaux du signal acoustique. Ceux-ci se manifesteront dans les paramètres acoustiques telles que la hauteur, l'intensité ou la distribution spectrale de l'énergie. Les changements dans l'articulation produisent des modifications plutôt à court terme et segmentales du signal acoustique mesurées par la fréquence et la précision des formants. Toutefois, ces modifications segmentales pourraient aussi susciter des changements à long terme lorsqu'elles se prolongent sur de longues périodes de la parole.

Ce concept d'effets " pull/push " en tant que deux types majeurs de déterminants des signaux vocaux nous paraît directement lié au modèle des multiples fonctions de Bühler. On pourrait prétendre que la fonction de symptôme, c'est-à-dire l'expression d'un état interne représente les effets " push ", alors que le symbole et l'appel seraient des aspects représentant les effets " pull ". Différents facteurs pourraient déterminer la nature de l'expression dans chaque cas. Et il pourrait y avoir un antagonisme entre effets " push " et " pull ", comme dans le cas par exemple où une excitation physiologique accrue " pousserait " la fréquence fondamentale à un niveau plus élevé, alors que les tentatives conscientes de contrôle la " tireraient " vers le bas, ce qui peut mener à la production de messages mixtes, voire contradictoires. S'il en est ainsi, il serait empiriquement important d'isoler ces deux déterminants. Recherche et théorisation futures dans ce domaine seront nécessaires pour différencier plus exactement entre ces multiples déterminants et multiples fonctions afin d'éviter de futiles controverses sur la " vraie nature " des vocalisations affectives.

3. Evaluation empirique des déterminants multiples

Le type de déterminant devrait avoir un effet important sur le codage, c'est-à-dire sur la relation entre le référent sous-jacent et les caractéristiques du signal. Dans une condition " push ", la tension des muscles augmenterait sous l'effet du stress en produisant une augmentation de la fréquence fondamentale de la voix (Fo). On s'attend à une covariation directe entre la quantité d'augmentation de tension musculaire, mesurée par électromyographie, et l'augmentation de la fréquence fondamentale. Dans ce modèle, la covariation attendue est continue (linéaire ou non-linéaire) entre les deux classes de variables.

Une alternative à ce modèle serait celui que nous appellerons le modèle configuration. Ce modèle est plus " linguistique " que le modèle covariation, lui-même plus psychologique par nature. Le modèle configuration soutient que pour réaliser un certain effet sur l'auditeur, l'individu utilise une combinaison particulière d'intonation, d'accent, de mots et de structures syntaxiques, par exemple une élévation du contour de l'intonation dans une interrogation et une diminution de cette intonation pour une question appelant une réponse de type oui/non. Il n'y a pas de dimensions variables ni continues dans les effets du modèle de configuration. Certaines classes de phénomènes doivent se coproduire pour susciter un effet. En terme d'effets " push " et " pull ", " push " est susceptible de suivre les règles du modèle de covariation, alors que " pull " suivrait plutôt les règles du modèle de configuration. Afin de mieux comprendre comment les processus de communication suivent soit le modèle de covariation soit celui de configuration, nous avons besoin de récolter davantage de connaissances sur les déterminants.

Scherer, Ladd et Silvermann [Sch84] ont mené deux études pour distinguer covariation et configuration. La première a utilisé un corpus de questions tirées d'une vaste étude sur les interactions entre fonctionnaires et citoyens. Celles-ci étaient homogènes dans leur structure mais variaient en terme de style pragmatique. Certaines questions étaient clairement des reproches, formulés comme des interrogations, alors que d'autres étaient des questions d'information purement factuelle. Trois techniques de filtrage ont été utilisées *a) low pass filtering, b)"random splicing" et c) "reversing"* (voir Sch85 pour une liste comparative des indices acoustiques

retenus par les techniques respectives). Les résultats démontrent que même lorsque le contenu des questions est rendu inintelligible, une grande part de la signification affective reste dans le signal acoustique. Le modèle de covariation se trouve confirmé dans son affirmation que les indices vocaux non-verbaux convoient l'affect de façon directe et indépendante du contexte.

Néanmoins, ces résultats n'apparaissent que pour les conditions de masquage dans lesquelles les paramètres de la qualité de la voix sont audibles, c'est-à-dire dans les techniques de *random splicing* et *reversing*. Dans ces deux conditions, le contour d'intonation de la phrase est perdu ou détruit. Ces caractéristiques ne joueraient donc pas un rôle essentiel dans la communication de la signification affective.

Evidemment, il y a des preuves empiriques contradictoires et contre-intuitives montrant que cette information est pertinente pour la communication de l'affect. Serait-ce que l'intonation suit les règles du modèle de configuration plutôt que celles du modèle de covariation ? Afin d'étudier cette question, avec le matériel de cette étude, nous avons divisé les questions en interrogations et en questions appelant des réponse de type oui/non et classifié les contours d'intonation en chute finale ou en augmentation finale. Les résultats montrent qu'il y a une forte interaction entre type de question et courbe d'intonation. Donc, on constate que certains paramètres acoustiques, telle que la qualité de la voix, opèrent selon les règles du modèle de covariation alors que d'autres, tel que le contour d'intonation, sont utilisés selon les règles du modèle de configuration.

Cet argument gagne davantage de sens si on le considère en terme d'approche psychobiologique de la communication. En effet, les paramètres qui montrent un degré remarquable de continuité phylogénétique, comme la nature différentielle de la phonation qui mène à des qualités de voix différentes, pourraient être plus proches du modèle de covariation directe avec les états physiologiques. A l'inverse, les paramètres qui ont été " domestiqués " avec le système de langage, telle que l'intonation, suivraient un modèle de configuration.

Afin de tester davantage ces notions, nous avons utilisé des techniques de synthèse et resynthétisation digitale de la parole pour évaluer l'influence d'indices prosodiques spécifiques et de la qualité de la voix sur l'impression de l'auditeur. Une telle approche évite évidemment le désavantage d'utiliser un corpus naturel, puisqu'elle permet un contrôle expérimental plus étroit des variables étudiées qui peuvent être manipulées indépendamment l'une de l'autre, tout en maintenant les autres paramètres acoustiques constants. Dans une série de travaux, nous avons employé cette technique pour varier systématiquement le contour d'intonation, la fréquence fondamentale (Fo), l'intensité, le rythme, l'accent, la structure et d'autres paramètres. [Ber88; Lad85; Tol88].

Trois résultats essentiels ont été mis en évidence: d'abord, nous n'avons pas trouvé d'effet d'interaction, ce qui suggère que les variables acoustiques que nous avons étudiées fonctionnent largement indépendamment l'une de l'autre. Deuxièmement, dans ces études où nous avions plusieurs locuteurs et plusieurs phrases, nous n'avons en réalité trouvé aucune interaction entre ces facteurs et les paramètres acoustiques manipulés. Ceci nous encourage à penser que les effets peuvent être généralisés à une grande variété de locuteurs et de phrases. Enfin, la fréquence fondamentale se présente comme celle qui a de loin le plus puissant effet sur le juge, particulièrement sur ses attributions de l'excitation. De plus, ces attributions semblaient être une fonction continue des changements dans la variabilité de la fréquence fondamentale, puisque les augmentations d'excitation rapportées étaient linéaires avec les augmentations de la fréquence fondamentale. Les résultats pour les contours d'intonation et la qualité de la voix sont complexes et demandent encore à être précisés par des travaux ultérieurs. En ce qui concerne les contours d'intonation, la difficulté pourrait être due en partie au rôle important joué par le modèle de configuration pour cette variable; en conséquence, nous considérons que la distinction entre règles de configuration et règles de covariation pourrait être très utile pour la compréhension de la communication de l'affect par la parole et il nous apparaît utile de continuer ce type de recherches avec l'aide des techniques modernes de manipulation des signaux digitaux.

4. Les profils acoustiques de l'émotion

Les travaux mentionnés ci-dessus ont trait à des états affectifs d'une intensité relativement basse, proches de ceux que nous rencontrons dans nos interactions sociales normales. Les études sur l'encodage, c'est-à-dire concernant l'effet des changements physiologiques dans l'état du locuteur sur ses productions vocales, ont été menées en induisant expérimentalement un état émotionnel spécifique ou en utilisant des prestations d'acteurs simulant un état émotionnel afin d'établir, par des méthodes de traitement digital, les conséquences de ces états affectifs sur la voix et la parole [Sch84; Sch91; Wal86]. Pittam et Scherer [Pitt93] ont résumé l'état de la littérature à ce jour comme suit :

COLERE : elle semble être, généralement, caractérisée par une augmentation de la moyenne de la F0, de la moyenne de l'énergie, de l'énergie haute fréquence ainsi que par des contours de la F0 descendante. La vitesse d'articulation habituellement augmente. Certaines études, qui ont mesuré la colère " chaude " ou rage (la plupart des études ne définissent pas si elles étudient la rage ou la colère " froide " plutôt synonyme d'irritation) ont également montré une augmentation de la variabilité de la F0 et du champ de la F0 à travers les phrases encodées. Les études qui n'ont pas trouvé de telles caractéristiques pourraient avoir mesuré en fait la colère froide.

PEUR : il y a un accord important sur les paramètres acoustiques associés à la peur. On s'attend à ce qu'un niveau d'excitation soit associé à cette émotion et ceci semble appuyé par l'augmentation de la moyenne de la F0, du champ de la F0 et de l'énergie haute fréquence généralement observée. La vitesse d'articulation est plus élevée. Une augmentation dans la moyenne de la F0 a également été trouvée dans des formes plus douces de

l'émotion, comme dans l'inquiétude ou l'anxiété.

TRISTESSE : comme pour la peur, on rencontre une bonne convergence entre les études qui ont pris en compte cette émotion. Une diminution de la F0, de l'étendue de la F0 et de la moyenne de l'énergie est en principe observée, de même qu'une tendance des contours de la F0 descendante. L'énergie haute fréquence et la vitesse d'articulation diminuent. La plupart des recherches ont toutefois étudié des formes plus tranquilles de cette émotion plutôt que des variantes très aiguës comme le désespoir. Ce dernier pourrait être caractérisé par une augmentation de la F0 et de l'énergie.

JOIE : c'est l'une des rares émotions positives étudiées, le plus souvent sous sa forme " forte " comme l'allégresse plutôt que sous ses types atténués comme le plaisir ou le contentement. Consistant avec un niveau élevé d'excitation auquel on pouvait s'attendre, nous avons trouvé une forte convergence des recherches sur une augmentation de la moyenne de la F0, de l'étendue de la F0 et de la variabilité de la F0, de même que de la moyenne d'énergie. Certains travaux ont mis de plus en évidence une augmentation de l'énergie haute fréquence et de la vitesse d'articulation.

DEGOUT: les résultats pour le dégoût tendent à être inconsistants au travers des études. Le peu d'entre elles ayant inclus cette émotion varient dans leur procédure d'encodage pour la mesure du dégoût, s'appuyant tantôt sur des films désagréables, tantôt sur des prestations d'acteurs simulant l'émotion. Les études utilisant des films ont relevé une augmentation de la moyenne de la F0, alors que celles utilisant des simulations d'acteurs ont indiqué l'inverse, à savoir une diminution de la F0. Cette inconsistance trouve écho dans la littérature sur le décodage.

Les travaux empiriques ne se sont pas limités à l'étude d'émotions primaires mais se sont également progressivement intéressés aux effets du stress ou encore aux états affectifs pathologiques, telle la dépression ou la schizophrénie, qui semblent également s'accompagner de caractéristiques vocales spécifiques.

STRESS: une déviation de la F0 à partir de sa ligne de base paraît indiquer les changements de magnitude d'un stress émotionnel [Kur76; Sim80; Gri87]. Plus spécifiquement, un stress émotionnel ou des conditions de charges mentales intenses mènent à une élévation des valeurs de la F0. Des effets similaires ont été mentionnés concernant l'amplitude et la durée des phrases parlées: chez des sujets vivant un stress majeur, l'amplitude moyenne de l'enveloppe spectrale est généralement plus grande alors que la moyenne de durée des phrases est plus basse [Gri87]. Au contraire, Tolkmitt et Scherer [Tol86] ont trouvé que le parterre de la F0 est plus élevé lors de stress ou cognitif ou émotionnel chez des sujets très anxieux ou répresseurs, alors que chez des sujets peu anxieux ce changement de F0 est opposé. De plus, et indépendamment du type de stress, une modification de la précision des deux premiers formants a été observée, la direction de ces variations dépendrait toutefois des stratégies de coping utilisées par le locuteur.

DEPRESSION : les résultats les plus consistants concernent l'effet de la dépression sur l'intensité de la voix. Une étude menée par Ellgring et Scherer [Ell96], a pu confirmer un niveau de F0 élevé dans les phases dépressives aigues, amélioré après traitement (comme constaté par des recherches antérieures, p.e. Ekm76; Tol82), mais seulement pour des femmes dépressives, pas pour les hommes. Par contre, ces auteurs trouvent un ralentissement du débit chez les deux sexes.

Même si ces résultats semblent indiquer une différenciation acoustique relativement claire entre les émotions fondamentales, on ne peut cependant pas exclure que la majeure partie des différences soit due au simple facteur excitation – excitation sympathique élevée, qui est caractéristique de plusieurs émotions et qui serait responsable de l'augmentation de la F0, de l'énergie et de l'énergie spectrale haute fréquence. La question de savoir si l'expression vocale indique seulement une excitation sympathique, plutôt que des différences qualitatives entre émotions (comme trouvé dans le cas des expressions faciales prototypiques) a été une autre préoccupation importante dans le domaine.

Une étude récemment menée par notre groupe de recherche permet une avancée concernant la question [Ban96]. Douze acteurs professionnels ont été priés de jouer 14 émotions en variant intensité et valence (qualité). Au total, 224 prestations différentes ont été recueillies, et ont été présentées à des juges auxquels on demandait de décoder ou d'inférer la catégorie émotionnelle renvoyée par l'émetteur. Les résultats répliquent les recherches antérieures démontrant la capacité des juges à inférer des émotions exprimées vocalement avec une précision supérieure à celle qui pourrait être attribuée à la chance. Le tableau 1 présente les différences dans la précision de reconnaissance pour 14 émotions.

Tableau 1: Classification des représentations d'émotions vocales par des juges, *jack-knifing*, et analyse discriminante en pourcentage

	Emotions représentées														
Emotions reconnues	CCh	CFr	PPan	Anx	Dés	Tris	Allég	Bonh	Int	Ennu	Hon	Fier	Dég	Mép	*Sum*
Colère chaude	**78**	17	10		6		14							2	127
	69		13		25		6		6						119
	75		13		13		13		6						120
Colère froide	10	**34**	2	7	5	1	5	1	3	2	2	5	10	10	97
		44		7				6	13			13	19	6	108
	13	**50**	6						13	6		13	13		114
Peur panique			**36**	13	9	1	7				1		1		68
	6	19	**31**		13		6								75
	6	13	**50**		6		13								88
Anxiété			27	**42**	18	2	5	1	1		15	2	10	2	125
				60	6	6		13	6	19	25	25		6	166
				53	6	25		19		6	19	13			141
Désespoir			21	7	**47**	21	16	1		1	4		10		128
	6	13	19		**38**		13								89
		13	13		**50**		13								89
Tristesse				5	8	**52**		3		13	19	2	14	8	124
			6			**44**				19	13	6	6	6	100
		6				**63**					13	6	6	6	100
Allégresse		1			1		**38**	2				4			46
	13		19		19		**69**								120
	6		6		19		**63**								94
Bonheur		2		4	1	3	1	**52**	8	1	8	23	5		108
				20		13		**44**			19	13	6	6	121
		6						**50**	6		6			13	81
Intérêt		7	1	7		1	4	18	**75**	1	13	12	2	2	143
		13	6	7				13	**56**		6	19	6		126
				7				6	**50**		13	13		6	95
Ennui		4		1		5		4	1	**76**	4	2	2	4	103
		6				19				**38**	6		6	13	88
		6		13	6					**56**	6		6	13	106
Honte			1	5	2	9	1	3	1	1	**22**	2	10	2	59
								13			**13**		6		31
				7				6		6	**31**		6		56
Fierté	1	15	1	4		2	2	17	10	1	8	**43**	7	6	117
	6		6	7		6	6	6		6		**13**	25	19	100
		6	6	7		6		13	13		6	**56**	19		132
Dégoût	1	2	1	2	2		2		1		1		**15**	5	32
		6				6		6	13	6	13	6	**13**	6	75
			6	13		6			6	6			**50**	25	112
Mépris	11	18	1	4	3	3	7	1	1	6	4	6	15	**60**	140
						6			6	13	6	6	13	**38**	88
								6		19	6	6		**38**	75

Note: Les résultats de l'étude sur la reconnaissance sont présentés dans la première ligne de chaque cellule, les résultats du *jack-knifing* et de l'analyse discriminante dans la seconde et troisième ligne respectivement. Les cellules vides représentent les valeurs zéro.

Les différences mesurées de façon consistante dans la " reconnaissabilité " de diverses émotions furent également répliquées. Toutes les prestations des acteurs furent soumises à des analyses acoustiques digitales de façon à obtenir des profils des paramètres vocaux pour chaque émotion, en utilisant un grand nombre de variables acoustiques. Le tableau 2 présente les profils acoustiques pour les émotions les plus intéressantes.

Les données indiquent que les paramètres vocaux non seulement indexent le degré d'intensité typique pour les diverses émotions mais différencient également la valence ou les aspects qualitatifs. Elles suggèrent de plus qu'avec davantage de finesse dans la mesure acoustique, il pourrait être possible de déterminer des profils acoustiques stables pour la plupart des émotions, pourvu que les états émotionnels différenciés soient utilisés de façon appropriée. Des analyses discriminantes et *jack-knifing* ont été conduites afin de déterminer avec quel degré de réussite les 14 émotions pouvaient être différenciées sur la base des paramètres vocaux mesurés. Les résultats montrent de façon remarquable que les réponses correctes et les patterns de confusion reflètent étroitement ceux donnés par des juges. Les profils acoustiques de la tristesse et de la honte (Figure 1) illustrent comment deux émotions exprimées avec des charactéristiques acoustiques semblables sont hautement confondues, autant par des juges humains que par des classifications informatisées. On a pu également trouver qu'une bonne part de la variance dans les inférences des juges pouvait être prédite sur la base des mesures acoustiques en permettant une évaluation plus détaillée des indices acoustiques que les auditeurs utilisent dans l'inférence de l'émotion à partir de la voix.

Tableau 2: Profils vocaux de 14 émotions: moyennes et déviations standards de 29 paramètres acoustiques résiduels transformés en scores z (Sexe de l'Acteur et Identité de l'Acteur partialisés).

	CCh	*CFr*	*PPan*	*Anx*	*Dés*	*Tris*	*Allég*	*Bonh*	*Int*	*Ennu*	*Hon*	*Fier*	*Dég*	*Mép*
MF0	1.13	.16	1.23	-.58	.99	-.32	1.24	-.64	-.17	-.80	-.49	-.46	-.29	-1.03
P25F0	.92	.15	1.39	-.28	1.15	-.52	1.21	-.62	-.14	-.83	-.64	-.51	-.37	-.93
P75F0	1.13	.05	.91	-.83	.73	-.08	1.20	-.52	-.32	-.69	-.41	-.37	.00	-.85
SdF0	.50	-.10	-.63	-.86	-.73	.43	.21	.14	-.26	.07	.42	.07	.33	.35
MElog	1.19	.52	.84	-.37	1.00	-1.16	1.05	-.48	.19	-.54	-1.14	-.13	-.51	-.48
Dur Art	-.31	-.14	-.58	-.35	.32	1.04	.12	-.49	-.66	.70	.32	-.22	.08	.15
DurVo	-.45	.15	-.47	-.38	.07	1.25	-.34	-.45	-.42	.94	.20	-.06	.01	-.06
HammI	1.13	.29	.27	-.33	.90	-.43	.58	-.43	-.03	-.40	-.49	-.26	-.46	-.37
DO1000	-1.17	-.51	-.45	.16	-.72	1.32	-.66	.15	-.23	.70	.89	.04	.45	.05
PE500	-.55	-.58	-.12	.15	-.51	1.23	-.29	.32	-.30	.27	.51	-.09	-.17	.12
PE1000	-1.34	-.52	-.28	.53	-.59	.90	-.05	.39	.11	.44	.03	.35	-.11	.17
v-0.2K	-.43	-.40	-.33	.12	-.69	.66	-.34	.07	-.24	.85	.48	-.12	-.11	.49
v-0.3K	-.59	-.37	-.19	.89	-.33	.81	-.57	.23	-.27	.21	.59	.02	-.06	-.31
v-0.5K	-.13	-.36	.13	-.54	-.16	.85	.20	.25	-.10	-.19	.05	-.10	-.14	.19
v-0.6K	-.31	-.33	-.28	.62	-.13	-.05	.11	.11	.09	.16	-.18	.09	-.06	.19
v-0.8K	-.08	.62	-.14	-.18	.20	-.46	-.18	.00	.10	.13	-.21	.18	.30	-.29
v-1K	-.13	.18	.42	-.25	.18	-.66	.56	-.28	.49	-.32	-.46	.29	-.09	.06
v1-1.6K	1.46	.63	.30	-.44	.64	-.73	.09	-.56	.04	-.55	-.35	-.13	-.22	-.22
v1.6-5K	.86	.27	.17	-.44	.36	-.84	.04	-.07	-.21	-.26	.28	-.45	.33	-.05
v5-8K	-.33	-.12	.11	-.18	-.08	.38	-.31	-.21	-.24	.48	.18	-.23	.72	-.19
uv-0.25K	-.65	-.18	-.30	.33	-.56	1.04	-.65	.26	.39	.06	.46	.13	-.32	.00
uv-0.4K	.04	.52	.39	-.25	-.02	.14	-.30	.38	.10	-.36	-.31	-.13	.16	-.36
uv-0.5K	.75	-.17	.47	-.07	.01	.33	.14	-.15	.09	-.49	-.06	-.01	-.46	-.39
uv0.5-1K	-.12	.15	.27	-.23	.31	-.06	.05	.34	.02	-.44	-.34	.50	-.03	-.44
uv1-1.6K	.49	-.02	.14	-.08	.62	-.46	1.20	-.29	-.41	-.12	-.45	-.03	-.36	-.23
uv-2.5K	.62	.15	.04	-.04	.75	-.79	.43	.16	-.42	-.29	-.23	-.28	-.25	.14
uv2.5-4K	-.19	-.09	-.45	.28	-.72	-.11	-.61	-.06	.56	.19	.59	.17	.12	.32
uv4-5K	-.57	-.11	-.36	.17	-.52	.50	-.61	-.08	.15	.64	.22	-.22	.42	.39
uv5-8K	-.49	-.30	.11	-.04	-.40	.80	-.40	-.31	-.15	.51	.35	-.27	.37	.21

Légende: Fréquence fondamentale: MF0 = Moyenne, SdF0 = déviation standard, P25F0 = 25th pourcentile, P75F0 = 75th pourcentile; Energie: MElog = Moyenne; Taux de parole: DurArt = durée des périodes d'articulation, DurVo = durée des périodes de voix; Spectre moyen vocalisé: v-0.2K = 125-200 Hz, v-0.3K = 200-300 Hz, v-0.5K = 300-500 Hz, v-0.6K = 500-600 Hz, v-0.8K = 600-800 Hz, v-1K = 800-1000 Hz, v1-1.6K = 1000-1600 Hz, v1.6-5K = 1600-5000 Hz, v5-8K = 5000-8000 Hz; LMHamm = Indice Hammarberg; DO1000 = pente d'énergie spectrale au-dessus de 1000 Hz; PE500 = proportion d'énergie vocale jusqu'à 500 Hz; PE1000 = proportion d'énergie vocale jusqu'à 1000 Hz.; Spectre moyen non vocalisé: uv-0.25K = 125-250 Hz, uv-0.4K = 250-400 Hz, uv-0.5K = 400-500 Hz, uv0.5-1K = 500-1000 Hz, uv1-1.6K = 1000-1600 Hz, uv-2.5K = 1600-2500 Hz, uv2.5-4K = 2500-4000 Hz, uv2.5-4K = 4000-5000 Hz, uv5-8K = 5000-8000 Hz.

Les caractéristiques acoustiques d'une vocalisation émotionnelle pourraient refléter le pattern complet des processus d'évaluation cognitive qui a déterminé l'état émotionnel chez le locuteur. Cette évaluation des

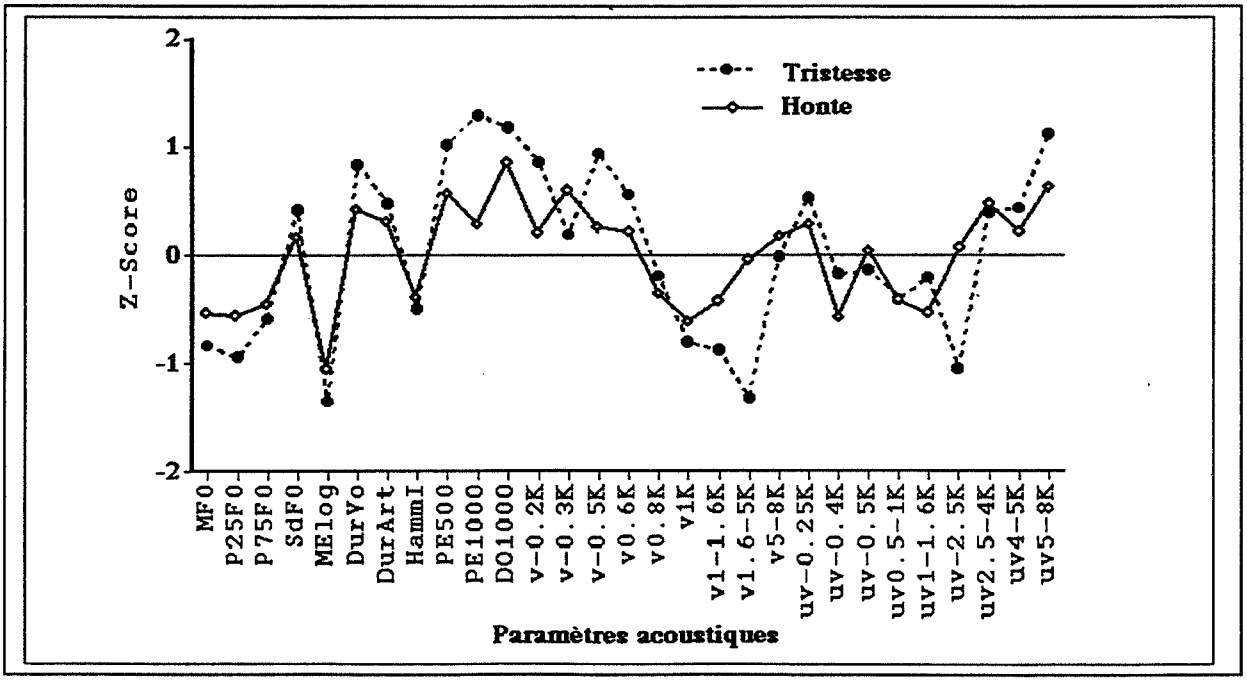


Figure 1 : Profils acoustiques de la tristesse et de la honte. Légende : voir tableau 2.

antécédents de l'émotion pourrait permettre à l'auditeur de reconstruire les caractéristiques majeures de l'événement et de ses effets sur le locuteur. Afin d'expliquer ce postulat, nous reprendrons certaines théorisations récentes sur l'émotion. Beaucoup de théoriciens du domaine de la psychologie de l'émotion semblent convaincus que la plupart des émotions humaines sont précédées par une évaluation cognitive de l'événement et de la situation, bien que les types de processus cognitifs puissent être relativement de bas niveau et non-conscients. Il est possible d'élaborer des prédictions sur comment on s'attendrait à ce que les caractéristiques phonatoires principales varient selon les critères d'évaluation des antécédents de l'émotion. Les données obtenues à partir de prestations d'acteurs furent utilisées pour tester ces prédictions théoriques sur le patterning vocal, basé sur le modèle des processus composant de l'émotion [Sch74]. Bien que la plupart des hypothèses soient supportées, certaines ont besoin d'être révisées sur la base des preuves empiriques accumulées [Ban96] .

5. Conclusion

La revue précédente de recherche sur l'effet de l'émotion sur la voix et la parole montre qu'une connaissance suffisante a été acquise pour commencer à essayer de transférer une partie de ces questions appliquées dans la technologie de la parole, en particulier la vérification du locuteur et la reconnaissance et synthèse de la parole. Evidemment, comme toujours, des recherches supplémentaires sont nécessaires, surtout depuis que la plupart de ce qui existe dans la littérature est limité aux langues anglaises et allemandes, et accompli par un nombre limité d'équipes de recherche. Toutefois, une grande part de cette recherche future pourrait être menée dans des cadres appliqués, abandonnant ainsi la distinction vaine entre recherche de base et recherche appliquée. Plus que dans beaucoup d'autres domaines, la recherche sur les effets du stress et de l'émotion sur la parole peut être menée avec profit dans le cadre d'un agenda de recherche ayant des buts appliqués clairement définis. Un exemple est un projet en cours sur la vérification du locuteur, menée par notre groupe (partiellement en collaboration avec d'autres laboratoires européens dans le projet VeriVox; Kar98]. Le but est d'entraîner des algorithmes de vérification du locuteur sur un assortiment plus large d'énonciations d'entraînement, contenant des échantillons obtenus sous différentes conditions émotionnelles. On s'attend à ce que l'élaboration de modèles qui définissent l'espace du locuteur quant à ses extensions potentielles dues aux effets émotionnels et attitudinaux, aidera à augmenter la robustesse (solidité) générale des algorithmes. En poursuivant ce but appliqué, l'étude produit également d'importants résultats de recherche de base sur la nature des signaux acoustiques qui différencient les états émotionnels et leur variabilité à travers les locuteurs, ainsi que beaucoup d'autres éclaircissements importantes.

L' exigence majeure pour une synergie prospère dans les efforts des ingénieurs tentant de perfectionner différents outils dans la technologie du langage, et des chercheurs provenant de disciplines contribuant à la science de la parole (telles que la phonétique et la psychologie), est l'abandon de la confiance exclusive dans la fiction selon laquelle les algorithmes statistiques améliorés s'occuperont des différences individuelles et des changements dans l'état du locuteur. Nous espérons que cette contribution puisse aider à fournir d'autres évidences que ceci consiste à prendre ses désirs pour des réalités plutôt que d'avoir des attentes réalistes. Comme il est montré au début de cette contribution, l'expérience pratique avec cette approche a été moins concluante jusqu'à maintenant, les critères de performance et d'acceptation étant quelque peu mal ciblés. De plus, nous avons tenté de montrer dans le reste de l'article, que des raisons théoriques suggèrent la nécessité de construire une compétence obtenue par une meilleure compréhension des mécanismes sous-jacents dans les algorithmes. Alors qu'une approche basée exclusivement sur les algorithmes d'ingénierie peut fonctionner dans certains domaines, il semble peu probable que la parole, en tant que système complexe de signaux combinant à la fois les systèmes de signification configurationnels et covariationnels, soit l'un d'entre eux.

De toute évidence, ce serait un grand ordre pour les ingénieurs de la parole de creuser dans les complexités de ce qui est déjà un domaine interdisciplinaire exigeant une compétence psychologique, physiologique et phonétique. Quoi qu'il en soit, il y a des exemples de collaboration fructueuse entre les ingénieurs et les scientifiques de la parole de différents milieux. Il nous semble que c'est ici la direction à prendre et nous désirons clore cette contribution par un appel à une collaboration accrue entre les différents groupes de chercheurs et de praticiens dans la communauté de la technologie de la parole.

Bibliographie

[Ban96] Banse, R. & Scherer, K. R. (1996), Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614-636.

[Ber88] Bergmann, G., Goldbeck, T., & Scherer, K. R. (1988). Emotionale Eindruckswirkung von prosodischen Sprechmerkmalen. *Zeitschrift für experimentelle und angewandte Psychologie*, 35, 167-200.

[Büh34] Bühler, K. (1934). *Sprachttheorie*, Jena: Fischer (new edition 1984).

[Dod98] Doddington, G. R. (1998). Speaker Recognition Evaluation Methodology. Proceedings of IEEE/ESCA RLA2C Workshop, Avignon.

[Ekm76] Ekman, P., Friesen, W. V., & Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, 16, 23-27.

[Ell96] Ellgring, H. & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20 (2), 83-110.

[Fri77] Frick, R. W. (1977). Prophylaxia in psychosomatic gynecology, as seen by a psychologist. *Psychotherapie und medizinische*

Psychologie, 27 (2), 53-57.

[Fur94] Furui, S. (1994). An overview of speaker recognition technology, Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1-9.

[Gri87] Griffin, G. R. & Williams, C. E. (1987). The Effects of Different Levels of Task Complexity on Three Vocal Measures. *Aviation, Space, and Environmental Medicine*, 58, 1165-1170.

[Joh96] Johnstone (1996). Emotional speech elicited using computer games, Proceedings of the 4th International Conference on Spoken Language Processing.

[Kar98] Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., & Scherer, K. R. (1998). Speaker verification with elicited speaking-styles in the Verivox project. Proceedings of IEEE/ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon.

[Kur76] Kuroda, I., Fujiwara, O., Okamura, N. & Utsuki, N., (1976). Method for Determining Pilot Stress Through Analysis of Voice Communication. *Aviation, Space, and Environmental Medicine*, 47, 528-533.

[Lad85] Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G. & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78, 435-444.

[Mar84] Marler, P. (1984). Animal communication: Affect or Cognition ? In K. R. Scherer & P. Eckman (Eds.), *Approaches to emotion* (pp. 345-368). Hillsdale, N. J. : Erlbaum

[Murr88] Murray, I. R., Arnott, J. L., & Newell, A. F. (1988). HAMLET-Simulating emotion in synthetic speech. Proceedings of Speech '88, The 7th FASE Symposium, Edinburgh, 4, 1217-1223.

[Pitt93] Pittam, J. & Scherer, K.R. (1993). Vocal Expression and Communication of Emotion. In M. Lewis and J. M. Haviland (Eds.) *Handbook of Emotions*. New York:Guilford Press.

[Sch74] Scherer, K. R. (1974a). Voice quality analysis of American and German speakers. *Journal of Psycholinguistic Research*, 3, 281-290.

[Sch77] Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilisation in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331-346.

[Sch81] Scherer, K. R. (1981). Speech and emotional states. In J. Darby (Ed.), *Speech evaluation in psychiatry*, 189-220. New York: Grune & Stratton.

[Sch84] Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76, 1346-1356.

[Sch85] Scherer, K. R., Feldstein, S., Bond, R. N., & Rosenthal, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research*, 14, 409-425.

[Sch86] Scherer, K. R. (1986). Vocal Affect Expression: A review and a Model for Future Research. *Psychological Bulletin*, 99, 143-165.

[Sch88] Scherer, K. R., Kappas, A. (1988). Primate Vocal Expression of Affective State. In D. Todt, P. Goedeking, & D. Symmes (Eds.), *Primate Vocal Communication* , 171-194. Berlin: Springer-Verlag.

[Sch90] Scherer, K. R., & Wallbott, H. G. (1990). Ausdruck von Emotionen. In K. R. Scherer (Ed.) *Enzyklopaedie der Psychologie*. Band C/IV/3 Psychologie der Emotion, 345-422. Göttingen: Hogrefe.

[Sch91] Scherer, K.R., Banse, R., Wallbott, H.G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, 123-148.

[Sch95] Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3), 235-248.

[Sim80] Simonov, P. V., Frolov, M. V. & Ivanov, E. A., (1980) Psychophysiological Monitoring of Operators Emotional Stress. *Aviation and Astronautics. Aviation, Space, and Environmental Medicine* 51, 46-50.

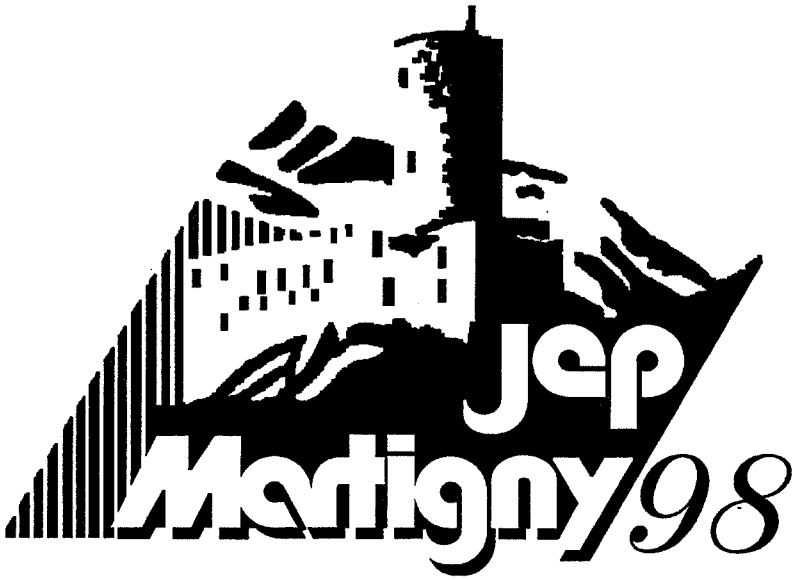
[Tol82] Tolkmitt, F., Helfrich, H., Standke, R., & Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of Communication Disorders*, 15, 209-222.

[Tol86] Tolkmitt, F. J. & Scherer, K. R. (1986). Effect of Experimentally Induced Stress on Vocal Parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 302-313.

[Tol88] Tolkmitt, F., Bergmann, T., Goldbeck, T, & Scherer, K. R. (1988). Experimental studies on vocal communication. In K. R. Scherer (Ed.), *Facets of emotion: Recent research*, 119-138. Hillsdale, NJ: Erlbaum.

[Wal86] Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51; 690-699.

Perception



Effet du renforcement des modulations temporelles en sous-bandes sur la reconnaissance de la parole: Résultats préliminaires

Frédéric APOUX[1], Frédéric BERTHOMMIER[2], Nicole BACRI[1], et Christian LORENZI[1]

[1] Institut de Psychologie, Laboratoire de Psychologie Expérimentale, URA CNRS 316,
Université René Descartes, Paris 5, 28, rue Serpente, 75006 Paris
Tél: +33/01 40 51 98 48 - Fax: +33/01 40 51 70 85
e-mail: christian.lorenzi@psycho.univ-paris5.fr
[2] Institut de Communication Parlée, UPRESA CNRS 5009,
INPG, 46, avenue Félix Viallet, 38031 Grenoble Cedex, France
Tél: + 33/04 76 57 48 28 - Fax: +33/04 76 57 47 10.
e-mail: bertho@icp.inpg.fr

Abstract

This study investigated the effects of enhancing the gross time-amplitude variations of speech stimuli by applying an expansive nonlinearity to the low-frequency temporal modulations of VCV stimuli presented in quiet or against a steady background noise. Temporal envelopes were extracted from broad spectral bands and magnified. All stimuli were then spectrally degraded, and so forced listeners to identify speech using primarily temporal cues. Preliminary results obtained with 4 normal-hearing listeners show no effects of envelope enhancement when speech is presented in quiet. The results also show consistent improvements in performance of 6-14% for the enhanced speech in noise. This suggests that the reduction in the modulation depth of the speech envelope caused by noise or reverberation could be efficiently compensated by enhancing low-frequency temporal modulations.

1. Introduction

L'enveloppe temporelle des signaux de parole présente des modulations d'amplitude de basse fréquence. Les modulations d'environ 3-4 Hz semblent les plus importantes [Hou85]. Deux groupes d'études ont montré que la présence de ces basses fréquences de modulation est une condition nécessaire et suffisante pour une bonne compréhension de la parole. Le premier groupe d'études a démontré que des signaux de parole dégradés spectralement sont compréhensibles tant que les modulations d'amplitude de basse fréquence sont préservées [Tas87 ; Sha95]. Le second groupe d'études a montré qu'une dégradation de ces modulations de basse fréquence causée par un lissage des modulations, la présence de bruit ou de réverbération, réduit l'intelligibilité [Duq80 ; Dru94a ; Dru94b ; Dru95 ; Hou85].

Le but de notre recherche est d'étendre nos connaissances relatives à l'importance de l'enveloppe temporelle de la parole en étudiant l'effet du renforcement des modulations temporelles sur l'identification de signaux de parole. Cet effet a été testé sur des stimuli dont l'enveloppe temporelle a été magnifiée par une fonction puissance. Cette approche a déjà été utilisée afin de simuler les effets du recrutement de sonie sur l'intelligibilité de la parole dans le bruit chez des sujets présentant une perte auditive d'origine cochléaire [Moo95 ; Nej97]. Les résultats de ces études montrent que le recrutement de sonie ne dégrade pas nécessairement la reconnaissance de la parole, mis à part le fait qu'il limite la dynamique du système auditif. Dans ces expériences, les effets du recrutement de sonie étaient mélangés à ceux produits par une élévation de seuil et une réduction de sélectivité fréquentielle (deux autres conséquences des lésions cochléaires). De plus, la structure fréquentielle des signaux était dégradée ou maintenue intacte, ne permettant pas de mesurer en isolation le rôle de l'enveloppe temporelle dans la compréhension de la parole.

Dans l'expérience rapportée ci-dessous, les stimuli étaient présentés à un niveau confortable d'écoute, et la structure spectrale fine des stimuli de parole était remplacée par celle d'un bruit blanc. Cette méthode, utilisée précédemment par Shannon et al. [Sha95], oblige ainsi les sujets à identifier la parole principalement sur la base d'indices temporels de basse fréquence (0-500 Hz).

2. Matériel et stimuli

Les stimuli étaient construits à partir d'une série de signaux de parole dits "non-traités". Ces derniers consistaient en 3 prononciations de chacun des 16 logatomes de type /aCa/ (C=/p,t,k,b,d,g,f,s,ʃ,m,n,r,l,v,z,ʒ/) lus par une locutrice française dans le silence. Chaque série de stimuli était présentée dans le silence et dans le bruit. Les stimuli de la condition "bruit" étaient construits en ajoutant à chaque logatome un bruit blanc stationnaire avec un rapport signal sur bruit (RSB) de 0 dB.

Chaque signal "non traité" était digitalisé à l'aide d'un convertisseur 16-bit A/D à une fréquence d'échantillonnage de 44.1 kHz, puis filtré en passe bas (filtre butterworth du 1er ordre, fréquence de coupure : 5kHz, -6 dB/oct.). Le signal était ensuite divisé en 4 bandes de fréquences (filtres elliptiques IIR du 3ème ordre): B1 : 20-800 Hz, B2 : 800-1500 Hz, B3 : 1500-2500 Hz et B4 : 2500-5000 Hz. Au sein de chaque bande, l'enveloppe était extraite par rectification mono-alternance et filtrage passe bas (filtre butterworth du 1er ordre, fréquence de coupure : 500 Hz, -6 dB/oct.). Les enveloppes obtenues dans chaque bande étaient ensuite centrées et normalisées par la puissance rms de l'enveloppe. Dans la condition "renforcée", l'enveloppe était élevée au carré; elle était laissée intacte dans la condition "non-renforcée". Chaque enveloppe était ensuite utilisée pour moduler un bruit blanc. Le bruit modulé résultant était filtré en utilisant les mêmes filtres passe-bandes que ceux du filtrage initial. Les bruits modulés filtrés étaient enfin combinés et filtrés en passe bas à 5000 Hz afin de produire le stimulus final dit "traité". Tous les stimuli étaient présentés de manière binaurale grâce à des écouteurs (Sennheiser HD 265). Afin d'évaluer l'effet des bandes spectrales sur la reconnaissance de la parole, une condition "nombre de bandes" a été introduite (B1, B1+B2, B1+B2+B3, B1+B2+B3+B4). Tous les stimuli étaient présentés à un niveau de 80 dBA pour un signal de parole continu.

3. Méthode

Les effets suivants étaient testés: (1) le nombre de bandes fréquentielles (1, 2, 3 et 4 bandes), (2) la présence du bruit masquant (conditions "silence" vs conditions "bruit"), et (3) le traitement de l'enveloppe (conditions "renforcées" vs "non-renforcées"). Les sujets étaient testés individuellement dans une pièce insonorisée. Le déroulement de l'expérience était contrôlé par un ordinateur (PC). Chaque session expérimentale était composée de 4 présentations successives des 48 stimuli VCV correspondants à une condition donnée, soit au total 192 stimuli VCV. Les sujets avaient pour consigne d'identifier chaque signal entendu au sein d'une liste présentée à l'écran. Pour chaque session (i.e., à partir des 192 présentations de stimuli VCV), le pourcentage de réponses correctes était calculé et une matrice de confusion était construite.

Tous les sujets débutaient l'expérience par deux conditions "non-traitées", une dans le silence, l'autre dans le bruit, ce dans un ordre contrebalancé entre les sujets. Les sujets étaient ensuite entraînés avec 4 sessions sur la condition [signal traité, 4-bandes, non-renforcé, dans le silence]. Les sujets passaient ensuite à la phase expérimentale, composée des 15 autres conditions et de la condition ayant servi d'entraînement. Les 16 conditions étaient présentées dans un ordre aléatoire, différent pour chaque sujet.

4. Sujets

Quatre sujets normo-entendants (FA, CL, CG, CD, âge moyen : 24 ans, écart-type : 3 ans) ont participé à l'expérience. Tous présentaient des seuils audiométriques tonaux inférieurs à 10 dB HL entre 125 et 8000 Hz.

5. Résultats

Dans chaque condition, le pourcentage de réponse correctes au hasard correspond à 6.25%. Les pourcentages de reconnaissance moyens pour les conditions non traitées dans le silence et dans le bruit sont respectivement de 99.6% et 89.2% avec des écart-types de 0.5% et 6.8%.

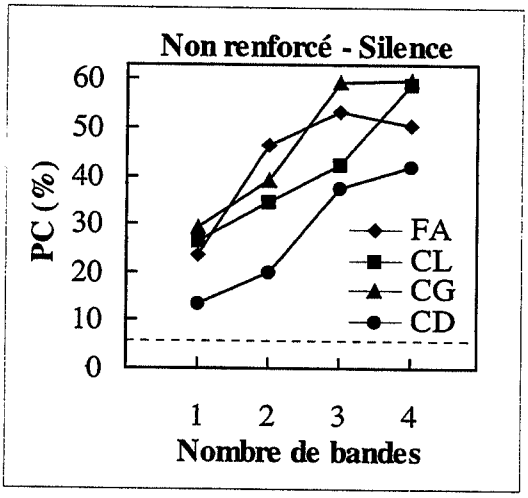


Figure 1 : Scores d'identification en fonction du nombre de bandes. Les stimuli de parole sont présentés dans le silence. L'enveloppe temporelle des signaux n'est pas renforcée. La ligne en pointillés indique le niveau de réponse au hasard (6.25%).

Les performances des sujets dans la condition [signal traité, non-renforcé, dans le silence] sont présentées dans la figure 1. Les performances des sujets sont toujours supérieures au hasard. Elles augmentent avec le nombre de bandes pour attendre un plafond situé entre 40% et 60% selon les sujets. Les niveaux de performance mesurés ici sont nettement inférieurs à ceux obtenus par Shannon et al. [Sha95]. Cette différence peut cependant s'expliquer par le plus grand nombre d'heures d'entraînement effectué par les sujets de Shannon et al. (8 à 10 heures). Ceci suggère que les performances de nos sujets n'étaient pas encore stabilisées au terme des 16 sessions expérimentales.

Les performances des sujets dans la condition [signal traité, renforcé, dans le silence] sont présentées dans la figure 2. Elles ne montrent pas d'effet de la surmodulation de l'enveloppe dans la condition 4 bandes. Toutefois, une dégradation des performances de 1 à 15% est observable dans les conditions 1, 2 et 3 bandes chez tous les sujets.

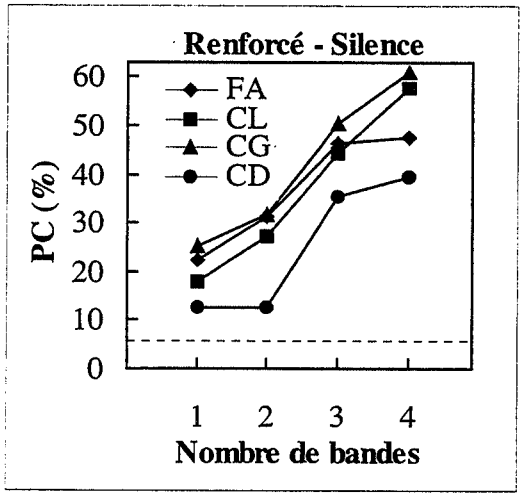


Figure 2 : Scores d'identification en fonction du nombre de bandes. Les stimuli de parole sont présentés dans le silence. L'enveloppe temporelle des signaux est renforcée.

Les figures 3 et 4 présentent les résultats obtenus dans le bruit respectivement sans et avec renforcement des modulations de l'enveloppe. Elles montrent toutes deux que l'introduction d'un bruit blanc stationnaire (+0 dB RSB) engendre une nette diminution des performances. L'effet du nombre de bandes est moindre et les performances ne dépassent pas 40% d'identification correcte. Ceci est compatible avec les résultats de Houtgast et Steeneken [Hou85], qui ont montré que les effets négatifs du bruit (ou de la réverbération) sur la parole peuvent être décrits en terme de réduction de la profondeur de modulation de l'enveloppe du signal dans chaque bande spectrale.

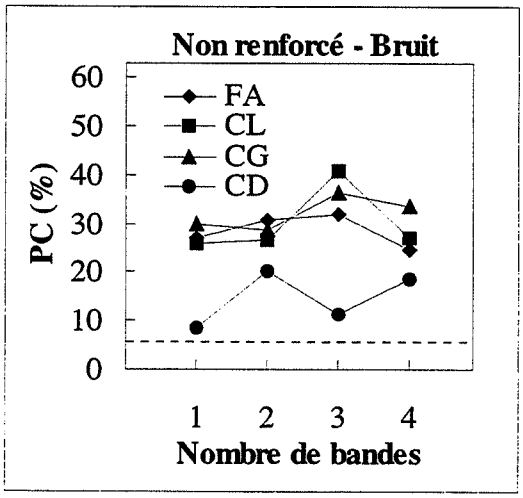


Figure 3 : Scores d'identification en fonction du nombre de bandes. Les stimuli de parole sont présentés dans le bruit (0 dB RSB). L'enveloppe temporelle des signaux n'est pas renforcée.

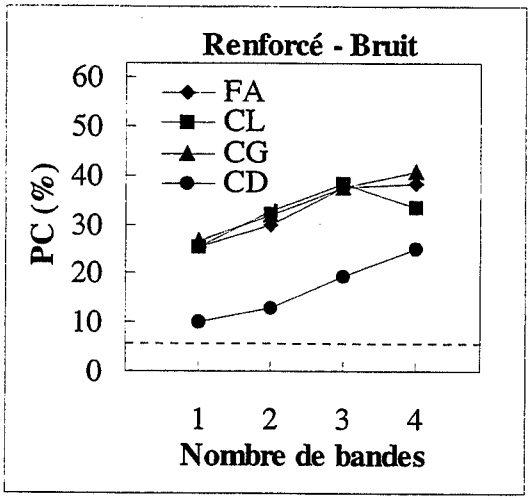


Figure 4 : Scores d'identification en fonction du nombre de bandes. Les stimuli de parole sont présentés dans le bruit (0 dB RSB). L'enveloppe temporelle des signaux est renforcée.

L'effet du nombre de bandes est plus évident en figure 4 (condition renforcée) qu'en figure 3 (condition non-renforcée). De plus, la figure 4 indique que la surmodulation produit un gain positif d'environ 6 à 14% dans la condition 4-bandes. Ces effets demandent à être confirmés avec plus de sujets. Toutefois, bien que très faibles, ils suggèrent que les effets négatifs du bruit pourraient être compensés efficacement par l'application d'une non-linéarité expansive à l'enveloppe temporelle du signal de parole.

Dans chaque condition expérimentale, une matrice de groupe a été également construite en ajoutant les matrices de confusion des 4 sujets. L'analyse détaillée de la matrice de confusion de groupe révèle que l'amélioration des performances de reconnaissance dans le bruit correspond principalement à une augmentation : (1) du pourcentage d'identification des occlusives orales /t,g/ et de l'occlusive nasale /n/, et (2) de la transmission du mode d'articulation (discrimination entre consonnes occlusives versus fricatives).

6. Conclusions

Dans cette expérience, nous avons préservé, dégradé ou renforcé les modulations temporelles lentes de signaux de parole, dans des conditions de dégradation spectrale sévère. Ces transformations étaient réalisées afin d'examiner l'importance et la robustesse des indices temporels dans la reconnaissance de la parole. Les résultats montrent que : (1) un bon score de reconnaissance de la parole peut être obtenu sur la base d'indices temporels, (2) une réduction de la profondeur des modulations d'amplitude lentes des signaux de parole dégrade le score de reconnaissance. Ces résultats confirment ceux obtenus dans des études antérieures [e.g., Dru94a ; Hou85 ; Sha95 ; Tas87]. Les résultats suggèrent également qu'une réduction de la profondeur des modulations temporelles causée par le bruit (ou la réverbération) peut être compensée en renforçant artificiellement l'enveloppe de parole. Ce résultat suggère des améliorations potentielles des méthodes de

reconnaissance automatique de la parole en sous-bandes [e.g., Bou97 ; Tib97]. Les troubles auditifs centraux produits par une lésion corticale [Alb74 ; Efr63] et les troubles d'acquisition du langage tels que la dysphasie ou la dyslexie développementale [Wri97] sont souvent associés à une mauvaise résolution temporelle auditive. L'effet positif du renforcement de l'enveloppe sur l'intelligibilité de la parole observé dans notre étude indique donc de nouvelles possibilités d'appareillage et de rééducation.

Bibliographie

[Alb74] Albert M.L. & Bear D. (1974), « Time to understand: A case study of word deafness with reference to the role of time in auditory comprehension », Brain, 97, pp. 373-384.

[Bou97] Bourlard H., & Dupont S. (1997), « Subband-based speech recognition », ICASSP, Vol. II, pp. 125-128.

[Dru94a] Drullman R., Festen J.M., & Plomp R. (1994), « Effect of temporal envelope smearing on speech reception », Journal of the Acoustical Society of America, 95, pp. 1053-1064.

[Dru94b] Drullman R., Festen J.M., & Plomp R. (1994), « Effect of reducing slow temporal modulations on speech reception », Journal of the Acoustical Society of America, 95, pp. 2670-2680.

[Dru95] Drullman R. (1995), « Temporal envelope and fine structure cues for speech intelligibility», Journal of the Acoustical Society of America, 97, pp. 585-592.

[Duq80] Duquesnoy A.J., & Plomp R. (1980), « Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis », Journal of the Acoustical Society of America, 68, pp. 537-544.

[Efr63] Efron R. (1963), « Temporal perception, aphasia and déjà vu », Brain, 86, pp.403-424.

[Hoc84] Hochmair E.S., & Hochmair-Desoyer, I.J. (1984), « Aspects of sound signal processing using the Vienna intra- and extracochlear implants », in R.A. Schindler & M.M. Merzenich (Eds), Cochlear implants, Raven Press, pp. 101-110.

[Hou85] Houtgast T., & Steeneken H.J.M. (1985), « A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria », Journal of the Acoustical Society of America, 77, pp. 1069-1077.

[Moo95] Moore B.C.J., Glasberg B.R., & Vickers D.A. (1995), « Simulation of the effects of loudness recruitment on the intelligibility of speech in noise », British Journal of Audiology, 29, pp. 131-143.

[Nej97] Nejime Y., & Moore B.C.J. (1997), « Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise », Journal of the Acoustical Society of America, 102, pp. 603-615.

[Sha95] Shannon R.V., Zeng F., Kamath V., Wygonski J., & Ekelid M. (1995), « Speech recognition with primarily temporal cues », Science, 270, pp. 303-304.

[Tas87] Van Tasell D.J.V., Soli S.D., Kirby V.M., & Widin G.P. (1987), « Speech waveform envelope cues for consonant recognition », Journal of the Acoustical Society of America, 77, pp. 1069-1077.

[Tib97] Tibrewala S., & Hermansky H. (1997), « Sub-band based recognition of noisy speech », ICASSP, Vol. II, pp. 1255-1258.

[Wri97] Wright B.A., Lombardino L.J., King W.M., Puranik C.S., Leonard C.M., & Merzenich M.M. (1997), « Deficits in auditory temporal and spectral resolution in langage-impaired children », Nature, 387, pp.176-178.

Perception des gestes de parole

René Carré* et Pierre L. Divenyi**

*)ENST, Département Signal, 46 rue Barrault, 75634 Paris cedex 13
**)Experimental Audiology Research, V.A. Medical Center, Martinez, CA 94553
carre@sig.enst.fr, pdivenyi@ucdavis.edu

Abstract

Perception of a synthetic French /aya/ sequence, generated by coproduction of two simultaneous gestures (tongue constriction and lip rounding), was tested in a series of experiments. The tokens consisted of different degrees of gesture reduction in which the target /y/ was not reached. The /ay/ trajectory in the plane F_1-F_2 was not rectilinear. The /aya/ transition is, therefore, an appropriate stimulus to study the role of the linearity of trajectories in speech perception. The findings point to the importance of the formant F_1-F_2 trajectory direction in vowel reduction and of the time dimension in phonetic distinctions.

1. Introduction

Dans le cadre de la phonologie articulatoire (Browman and Goldstein, 1986), l'élément de base n'est pas le phonème ou bien le trait distinctif mais plutôt le geste de parole, encore appelé geste articulatoire ou bien geste phonétique (Browman and Goldstein, 1989). Les caractéristiques de cet élément sont essentiellement dynamiques (et non statiques comme c'est le cas pour les traits distinctifs). La parole étant un phénomène acoustique, l'objet du geste de parole va être de provoquer des modulations acoustiques pouvant être perceptivement identifiées, en jouant sur des sources acoustiques ou bien sur la forme du conduit vocal pour en modifier les propriétés de résonance. La déformation spécifique du conduit vocal est réalisée au moyen des paramètres articulatoires (Saltzman, 1986). Plusieurs combinaisons différentes de ces paramètres permettent cette déformation et les meilleures solutions prennent en compte, en particulier, le contexte en appliquant un critère spécifique comme celui du moindre effort. Dans le cadre de la production des voyelles qui se distinguent essentiellement par les fréquences des deux premiers formants, on va s'intéresser à des gestes phonologiquement distinctifs de déformation du conduit vocal réalisant, pour une langue donnée, le passage d'une voyelle à une autre. Les gestes de déformation que l'on va utiliser dans la première partie de ce travail sont automatiquement obtenus à partir d'un critère d'efficacité acoustique (ou minimum d'énergie) de la déformation (Carré, et al., 1995). Ce critère conduit au modèle en régions distinctives (Mrayati, et al., 1988). On peut alors schématiser la production des transitions /ai/ et /ay/ selon la figure 1 (Carré and Mrayati, 1995). La transition /ai/ est obtenue, à partir de la configuration /a/, au moyen d'un seul geste transversal permettant de réaliser une constriction à l'avant du conduit vocal. A ce geste de constriction est automatiquement associé le développement d'une cavité à l'arrière du tube. Pour la transition /ay/, deux gestes sont nécessaires : au geste transversal précédent est associée une constriction aux lèvres. Les résultats acoustiques obtenus au moyen de cette modélisation qui peut paraître grossière, sont suffisants comme l'ont montré les synthèses correspondantes. Les transitions sont ainsi décrites le plus simplement possible en termes de gestes qui l'on qualifiera de phonologiquement distinctifs.

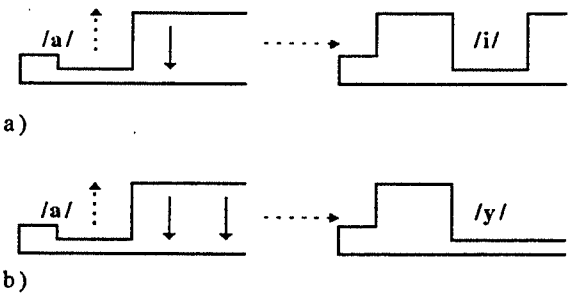


Figure 1. a) Production de la transition /ai/ au moyen d'un geste ; b) production de la transition /ay/ au moyen de deux gestes.

On a étudié certaines caractéristiques perceptives d'un geste (Carré, et al., 1994; Carré, et al., submitted) et montré que, dans une certaine limite, on peut, sans modifier le percept :

- modifier le mouvement dans le temps d'un geste,
- modifier le degré de synchronisation de deux gestes,
- 'réduire' un geste.

Nous nous proposons ici d'étudier la réduction vocalique lorsque deux gestes sont impliqués comme c'est le cas de la production de séquences /aya/.

2. Production d'ensembles /aya/. Trajectoires formantiques non rectilignes

On utilise le modèle en régions distinctives pour produire des séquences /aya/ selon le schéma de la figure 1b. Si le geste de commande de l'ouverture aux lèvres est strictement synchronisé avec le geste de réalisation de la

constriction avant, on obtient une trajectoire formantique dans le plan F1-F2 qui n'est pas rectiligne (figure 2). A titre de comparaison, nous avons aussi représenté sur cette figure la trajectoire /aia/ laquelle est pratiquement rectiligne. La trajectoire /aya/ pointe d'abord sur /i/ avant de se diriger vers /y/ (dans notre cas, le /y/ se distingue du /i/ par un 2ème formant plus bas ; on trouve aussi en français des situations où le /y/ se distingue du /i/ principalement par le 3ème formant). Cette aspect non rectiligne mérite discussion.

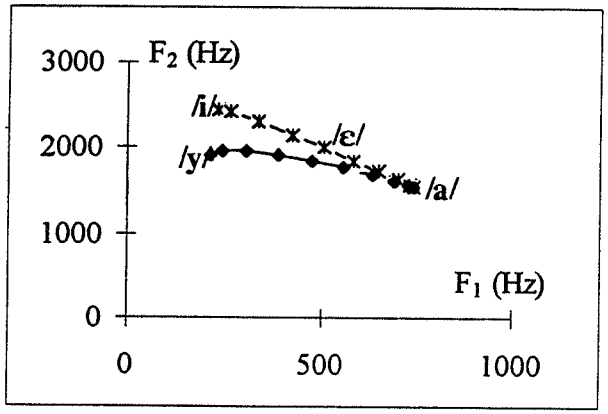


Figure 2. Evolution de la trajectoire formantique /aya/ dans le plan F1-F2. Elle n'est pas rectiligne. La trajectoire /aia/ est pratiquement rectiligne.

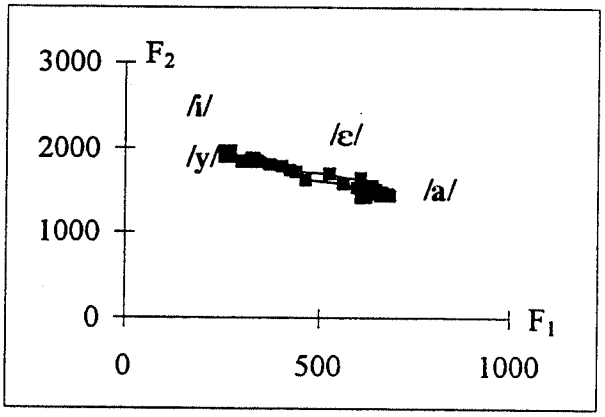


Figure 3. Trajectoire formantique de /aya/ en parole naturelle. Le grand carré à gauche représente une production isolée de /y/.

En effet, rappelons que :

- La trajectoire /aya/ obtenue en parole naturelle est rectiligne comme le montre la figure 3 (notons aussi que la cible statique correspondant à la production d'un /y/ soutenu et représentée par un grand carré est pratiquement atteinte) ;
- Des tests de perception ont montré que les trajectoires rectilignes sont préférées aux trajectoires non rectilignes dans le cas de production de type V1-V2 (Pasdeloup and Jospa, 1996) ;
- Enfin, en parole naturelle, on note généralement une anticipation de la commande labiale lors de la production des voyelles labiales. Cette anticipation permettrait effectivement de linéariser la trajectoire obtenue par le modèle DRM.

Au lieu de chercher à linéariser (par anticipation labiale) la trajectoire formantique, nous avons préféré la conserver telle quelle et étudier ses caractéristiques perceptives. Pour cela, nous avons produit 7 ensembles /aya/, dont certains ont des caractéristiques 'réduites'. Les trajectoires des 4 premiers atteignent effectivement la cible /y/ et y stationnent 30, 20, 10, 0 ms ; les trois derniers n'atteignent pas la cible et rebroussent chemin 10, 20, 30 ms avant d'atteindre la cible. La durée de la première voyelle est de 100 ms, les transitions de type cosinus ont une durée de 100 ms et la durée de la troisième voyelle est de 150 ms. La figure 4 montre les évolutions formantiques pour les 7 signaux.

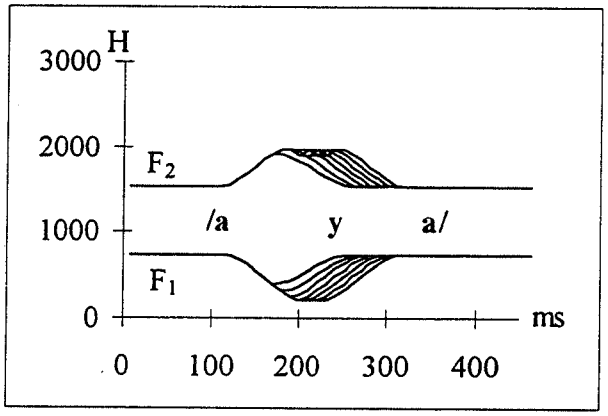


Figure 4. Transitions formantiques pour la production de /aya/.

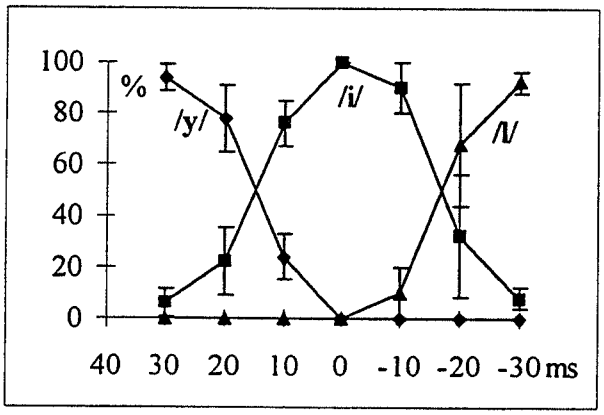


Figure 5. Résultats du test de perception.

Chacun des signaux est présenté 10 fois aléatoirement au cours d'une session. Cinq sessions ont été organisées. L'auditeur doit identifier la voyelle centrale parmi /y, i, ɛ, ø, l/. La figure 5 donne le résultat des tests d'identification pour un même auditeur. On note que :

- Il faut rester au moins 30 ms sur la cible /y/ pour qu'elle soit perçue comme telle ;
- Bien que la cible acoustiquement atteinte durant 10 ms soit /y/, le percept est la voyelle /i/. La forme de

la trajectoire dans le plan F1-F2 est donc importante. On doit envisager un mécanisme de mesure de la direction moyenne d'une trajectoire pour expliquer ce résultat.

- Ni la voyelle /ø/, ni la voyelle /ɛ/ ne sont perçues pour faire place au /l/.
- On peut difficilement parler de phénomène de réduction puisque le percept /y/ n'est jamais obtenu dans un contexte de cible acoustique non atteinte.

3. Production d'ensembles /aya/. Trajectoires formantiques rectilignes

On a reproduit l'expérience précédente en conservant les caractéristiques des cibles mais en utilisant une trajectoire formantique rectiligne (figure 6).

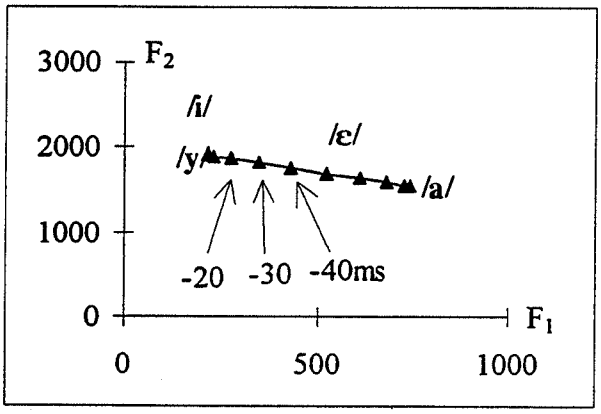


Figure 6. /aya/ avec trajectoire formantique rectiligne.

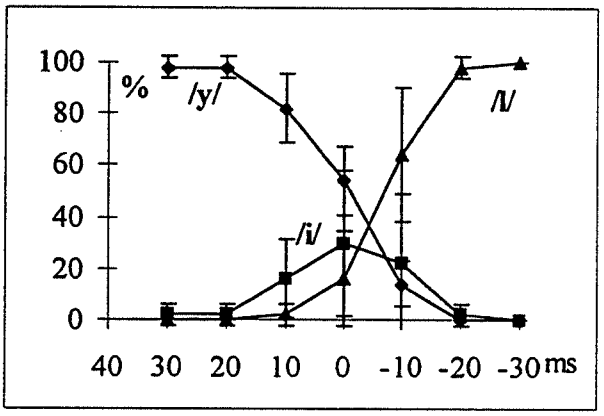


Figure 7. Résultat du test de perception dans le cas de /aya/ avec une trajectoire formantique rectiligne.

Les résultats des tests d'identification sont représentés figure 7. On note, en particulier :

- que la voyelle /i/ est très peu perçue ;
- que le percept /y/ est obtenu alors que la cible /y/ est juste atteinte ;
- qu'il n'y a pas de phénomène de réduction pour la perception de /y/.

On a alors voulu vérifier si ce mécanisme de réduction existe en parole naturelle. Pour cela, on a produit l'ensemble /aya/ avec un débit rapide. On note effectivement une légère réduction (figure 8).

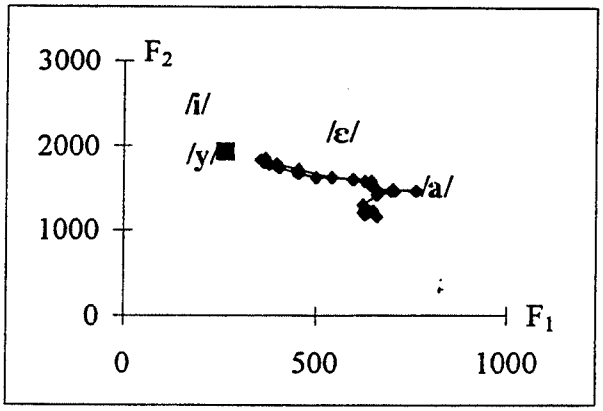


Figure 8. /aya/ en parole naturelle et débit rapide. Le grand carré indique la cible obtenue avec un /y/ soutenu. La trajectoire est plus ou moins rectiligne et on note un phénomène de réduction.

Par ailleurs, on a noté que la valeur du 3ème formant en parole naturelle pour /y/ est plus faible (2200 Hz) que dans nos simulations (2500 Hz). On a alors corrigé cette valeur et noté que, dans ce cas, le percept /i/ disparaît (figure 9) mais que le mécanisme de réduction n'apparaît toujours pas au profit de la perception de /l/.

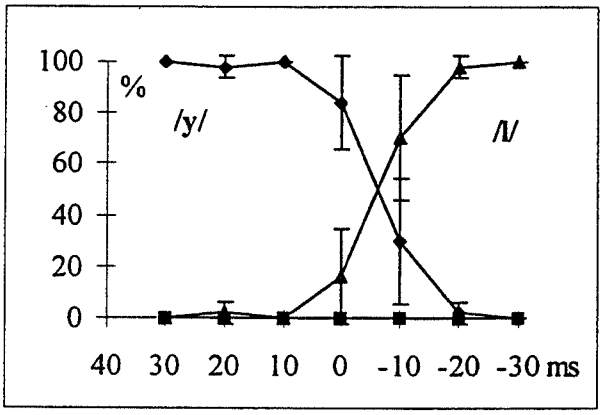


Figure 9. Le percept /i/ disparaît. Il n'apparaît pas de phénomène de réduction.

On se pose alors la question de l'origine du mécanisme de réduction constaté en parole naturelle puisque les trajectoires formantiques paraissent proches. On a alors noté des différences importantes dans le déroulement temporel des gestes et, en particulier, on a noté une stabilisation importante sur la voyelle perçue comme /y/ (figure 10).

On a aussi noté la proximité du percept /l/ qui pouvait limiter le développement d'un mécanisme de réduction. Cette proximité n'existant pas dans le cas de séquence /yay/, nous avons étudié l'existence d'un phénomène de réduction dans ce contexte. Comme pour les expériences précédentes, nous avons anticipé le retour vers la voyelle

finale de -20 à -60 ms. La cible /a/ n'est donc pas atteinte. La figure 11 montre clairement que la plage de perception de la voyelle centrale produite isolément est plus petite que celle de la voyelle prononcée en contexte. Dans ce cas, il y a bien réduction.

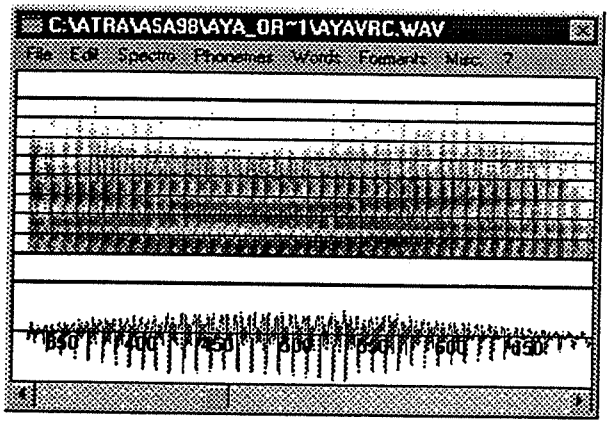


Figure 10. Sonagramme de /aya/ naturel prononcé à débit élevé. Les paramètres de la voyelle centrale perçue comme /y/ sont plus ou moins stables entre 420 et 520 ms. La trajectoire formantique correspondante est représentée figure 8.

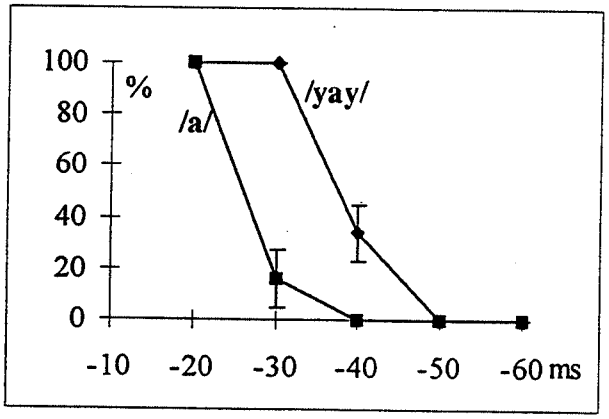


Figure 11. Résultats des tests de perception pour /yay/ et /a/.

Conclusions

Plusieurs conclusions découlent de nos résultats obtenus avec un seul auditeur. Mais notre expérience nous a déjà montré qu'une augmentation du nombre d'auditeurs augmente les écarts types uniquement parce que chaque auditeur a ses propres frontières entre percepts différents. La forme des trajectoires formantiques dans le plan F1-F2 joue un rôle important que la cible acoustique soit atteinte ou non. Les aspects temporels liés à la durée des transitions et à la durée de la cible intermédiaire dans le cas d'ensemble V1V2V1 jouent aussi un rôle important. Enfin, dans la mesure où la direction moyenne d'une trajectoire est conservée, on peut envisager une autonomie relative des deux gestes de parole mis en œuvre pour produire /aya/ mais cette autonomie paraît faible dans le cas de la réduction. Est-ce à dire que les deux gestes n'existent pas en tant que tels ? Des entités autonomes qui ont donc une existence propre se caractérisent par leurs degrés de liberté. Rappelons ce que dit Ohala (1992) en ce qui concerne les traits distinctifs : «...If features are linked by necessity, then they are not autonomous ». Ou bien n'y a-t-il pas ici aussi un principe d'incertitude : une imprécision pour atteindre une cible est compensée par une direction de trajectoire précise? Il reste donc bien des inconnues avant de pouvoir affirmer qu'il existe des gestes de parole.

Remerciements. Ce travail a reçu un soutien de l'OTAN.

Bibliographie

Browman, C. and Goldstein, L. (1986). "Towards an articulatory phonology," in *Phonology yearbook*, edited by C. Ewan and J. Anderson (Cambridge University Press, Cambridge) pp. 219-252.

Browman, C. P. and Goldstein, L. (1989). "Articulatory gestures as phonological units," Phonology 6, 201-252.

Carré, R., Ainsworth, W. A., Jospa, P., Maeda, S. and Pasdeloup, V. (**submitted**). "Perception of vowel-to-vowel transitions in terms of "speech gestures","

Carré, R., Chennoukh, S., Lindblom, B. and Divenyi, P. (1994). "On the perceptual characteristics of "speech gestures"," J. Acoust. Soc. Am. 96, S3326.

Carré, R., Lindblom, B. and MacNeilage, P. (1995). "Rôle de l'acoustique dans l'évolution du conduit vocal humain," Comptes Rendus de l'Académie des Sciences, Paris **t. 30, série IIb**, 471-476.

Carré, R. and Mrayati, M. (1995). "Vowel transitions, vowel systems, and the Distinctive Region Model," in *Levels in Speech Communication: Relations and Interactions*, edited by C. Sorin, J. Mariani, H. Méloni and J. Schoetgen (Elsevier, Amsterdam) pp. 73-89.

Mrayati, M., Carré, R. and Guérin, B. (1988). "Distinctive region and modes: A new theory of speech production," Speech Communication 7, 257-286.

Ohala, J. J. (1992). "The segment: primitive or derived?," in *Papers in Laboratory Phonology II: Gesture, segment, prosody*, edited by G. J. Docherty and D. R. Ladd (Cambridge University Press, Cambridge) pp. 166-183.

Pasdeloup, V. and Jospa, P. (1996). "Etude perceptive du déphasage entre les gestes labial et lingual en synthèse articulatoire," *Actes des XVIèmes Journées d'Etude sur la Parole*, (Avignon), pp. 19-22.

Saltzman, E. L. (1986). "Task dynamic coordination of the speech articulators: a preliminary model," in *Generation and modulation of action patterns*, edited by H. Heuer and C. Fromm (Springer-Verlag, New York) pp. 129-144.

Est-il absurde d'entendre /b/ dans "absurde"?

Céline CHEREAU, Pierre HALLE et Juan SEGUI

Université René Descartes, Laboratoire de Psychologie Expérimentale, URA 316 CNRS,
28, rue Serpente, 75006, Paris, France.
Tél.:01 40 51 98 65.

Abstract

Usually, surface and abstract phonological representations of speech segments are congruent. They become discrepant, however, as soon as assimilation rules apply. Such is the case for the labial plosive in "absurde", whose underlying and surface representations are /b/ and /p/, respectively. Listeners categorize the sounds modified through assimilation by relying on surface representations when they are excised from their speech context. When listeners process on-line meaningful utterances, however, they largely rely on abstract representations. Evidence for these two modes of speech perception ("phonetic" vs. "linguistic") is presented in two experiments. A phonetic gating task shows that listeners hear /p/, not /b/, in a truncated version of "absurde", /psyrd/. A phoneme monitoring task shows that listeners detect /b/ rather than /p/ in the unaltered word "absurde".

1. INTRODUCTION

Sous l'influence d'un son voisin, un son peut subir différents changements : deux sons contigus ont tendance à acquérir un ou plusieurs caractères communs. Ce sont ces changements qu'on regroupe sous le terme d'assimilation. Ainsi, le mot "absurde" contient le même son [p] que le mot "capsule", car le phonème sous-jacent /b/ est transformé en un phonème de surface /p/ par assimilation de voisement avec le /s/ suivant. En revanche, le mot "abjurer" contient le son [b]. Bien qu'on distingue théoriquement les sourdes douces (ici /p̬/) et les sonores fortes (ici /p /), il n'existe guère de différence acoustique entre ces deux sons [Rig67].

Rigault a montré que les auditeurs perçoivent le /t/ de /metsɛ̃/ ("médecin") selon sa forme phonologique de surface /t/ dans le mot tronqué /etsɛ̃/, mais le perçoivent selon sa forme linguistique sous-jacente /d/ dans le mot entier /metsɛ̃/. Ceci montre que les auditeurs croient entendre un phonème lorsque ils savent qu'il "faut" l'entendre, conformément à leurs représentations linguistiques abstraites.

Cependant, on ne sait pas comment les sons affectés par une assimilation sont traités en temps réel. Le premier objectif de l'étude présente est de vérifier que l'occlusive labiale de "absurde" est bien perçue comme [p] lorsqu'on l'isole de son contexte lexical. Dans une première expérience, nous utilisons pour cela une tâche de *gating* phonétique (cette tâche est une variante de la tâche de *gating* classique : les sujets ne doivent pas deviner un mot, mais transcrire ce qu'ils entendent). Le second objectif de cette étude est de savoir si, dans une situation de traitement en temps réel, le même son [p] active une représentation phonologique de surface /p/, ou une représentation sous-jacente plus abstraite /b/ (ce que Carton [Car74] appelle "forme linguistique"). A cette fin, une seconde expérience employant une tâche de détection de phonème, a été menée. Les sujets devaient détecter les phonèmes /p/ ou /b/ dans des mots comme "absurde" ("obtus", "abscisse", etc) et des mots comme "capsule" ("opticien", "lapsus", etc). Si les sons modifiés par l'assimilation de voisement sont perçus selon leur forme de surface, on s'attend à ce que les sujets détectent aussi facilement /p/ dans "absurde" que dans "capsule". En revanche, si les sujets perçoivent ces sons par le biais de leurs représentations linguistiques, on s'attend à ce que les sujets aient plus de difficulté à détecter /p/ dans "absurde" que dans "capsule". On s'attend aussi à une tendance massive à détecter /b/ à tort dans "absurde", beaucoup plus souvent que dans "capsule".

2. EXPERIENCE 1 : TACHE DE GATING

On sait que l'occlusive labiale dans "absurde" est le même son [p] que dans "capsule", c'est à dire une occlusive non-voisée. Mais il est possible que le voisement soit marqué par des indices subtils et difficilement quantifiables. Cette première expérience est destinée à contrôler cette possibilité. Nous avons utilisé une tâche de *gating* phonétique, dans laquelle les sujets devaient transcrire des items comme "absurde" et "abjurer", présentés de façon incrémentale mais privés du contexte gauche précédant l'occlusive (par exemple /psyrd/ et /bʒyre/ pour "absurde" et "abjurer"). Dans ces conditions de présentation, les sujets devraient faire appel à des représentations phonologiques de surface et pourraient identifier le voisement de ces fragments sans être perturbés par des représentations abstraites.

2.1 Méthode

2.1.1 Matériel

Le matériel est constitué de 16 mots, dont 8 de type "absurde" ou "obtus", et 8 de type "abjurer" ou "abdiquer". Une inspection visuelle des spectrogrammes de mots comme "absurde" et "abjurer" montre que, si l'occlusive labiale de "abjurer" est clairement prévoisée,

celle de "absurde" est totalement dépourvue de prévoisement, tout comme le /p/ de "capsule".
Pour les mots comme "absurde", le contexte gauche est supprimé jusqu'au silence d'occlusion. Pour les mots comme "abjurer", il est supprimé jusqu'au point de plus grande stabilité spectrale du prévoisement. La portion du prévoisement restant est atténuée par une rampe d'amplitude de 20 ms. Le premier fragment comporte ce qui est conservé avant le relâchement de l'occlusion et les 40 ms suivantes. La durée des sept fragments suivants augmente avec un incrément de 30 ms. Les stimuli sont présentés par blocs comprenant tous les fragments d'items correspondant au même numéro de gate (format de présentation "parallèle"). L'ordre des items est aléatoire, et change entre deux blocs successifs. Ceci permet d'éviter les problèmes de persévération dans les réponses, puisque les fragments présentés sont extraits d'items non prédictibles.

2.1.2 Procédure
La tâche des sujets était de transcrire ce qu'ils entendaient, et non de deviner un mot comme dans la tâche de *gating* classique. L'intervalle entre deux stimuli était de 4 secondes.

2.1.3 Sujets
Douze étudiants de Pyschologie ont participé à cette expérience. Tous étaient de langue maternelle française et ne présentaient aucun trouble auditif.

2.2 Résultats et discussion

Deux analyses de variance, par sujets et par items, ont été conduites séparément pour les deux types de mots.

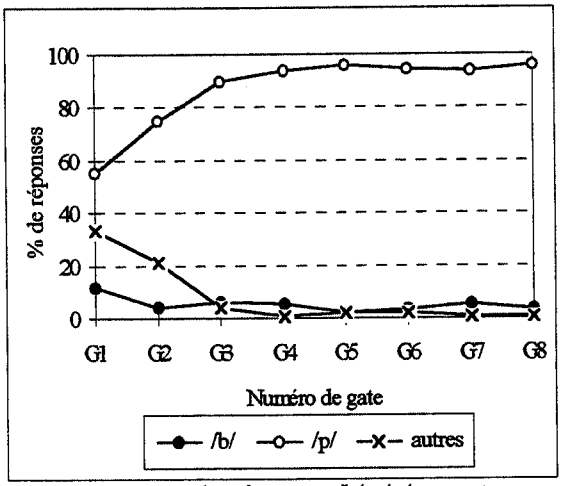


Figure 1 : taux de réponses /b/, /p/ et autres pour les items de type "absurde" /"obtus"

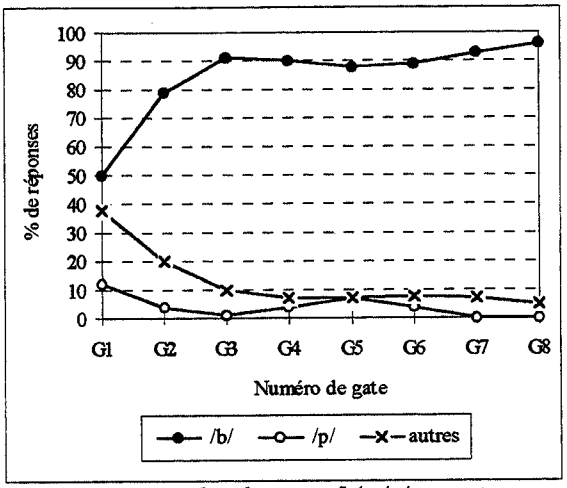


Figure 2 : taux de réponses /b/, /p/ et autres pour les items de type "abjurer" /"abdiquer".

Pour les items comme "capsule" ou "obtus", le pourcentage de réponses /p/ atteint 96% et celui des réponses /b/ 3% environ, au dernier *gate*. Dans les deux cas, les réponses des sujets sont donc massivement "correctes". De plus, les pourcentages de réponses /p/ pour les mots comme "absurde" ne sont pas différents des pourcentages de réponses /b/ pour les mots comme "abjurer".
Ces résultats permettent de conclure que, dans une situation de présentation incrémentale, les réponses des sujets reflètent une perception phonétique de bas niveau.
Une objection possible à cette conclusion serait de dire que les réponses reflètent un biais lexical induit par les contraintes phonotactiques du Français. En effet, si les séquences /ps/ ou /pt/ existent en début de mot (au moins dans quelques mots, comme "psychologie", "ptérodactyle"...), les séquences /bs/ ou /bt/ sont phonotactiquement illégales en début de mot. Ceci aurait donc pu favoriser les réponses /p/. Cependant les séquences /bʒ/ et /bd/ sont tout aussi illégales en début de mot que les séquences /bs/ et /bt/. Or les réponses /b/ dans les fragments venant de "abjurer" sont aussi nombreuses que les réponses /p/ pour les fragments venant de "absurde". Les réponses des sujets ne reflètent donc pas un biais phonotactique ou lexical. Il a de plus déjà été montré que la tâche de *gating* phonétique fait plutôt intervenir des représentations phonétiques de bas niveau. Par exemple, les sujets transcrivent correctement les séquences sonores initiales /dl/ ou /tl/, illégales en début de mot [Hal98]. On est donc en mesure d'affirmer que l'occlusive présente dans les mots comme "absurde"ou "obtus" est bien perçue /p/ lorsqu'elle est isolée de son contexte lexical.

3. EXPERIENCE 2 : TACHE DE DETECTION DE PHONEME

Les auditeurs entendent bien [p] dans "absurde" lorsque ce son est sorti de son contexte lexical. Que se passe-t-il lors du traitement en temps réel des mots comme "absurde", lorsque tout le contexte lexical est conservé?
Nous utilisons pour le savoir une tâche de détection de phonème généralisée, dans laquelle les sujets doivent détecter /p/ ou /b/ dans des mots tels que "absurde" et "capsule". Dans cette situation, si les sujets utilisent une représentation phonologique abstraite, ils devraient détecter /p/ plus difficilement dans "absurde" que dans "capsule", et détecter /b/ plus souvent que /p/ dans "absurde". Il est aussi possible que l'accès à une représentation linguistique abstraite soit plus lent que l'accès à une représentation phonologique de surface.

3.1 Méthode

3.1.1 Stimuli

Trente deux mots contiennent le phonème /p/, représenté pour la moitié d'entre eux par la lettre 'b' suivie de /s/ ou /t/ ("absurde" ou "obtus") et pour l'autre moitié par la lettre 'p' ("capsule" ou "opticien"). Il y a aussi 64 items de remplissage contenant le son /b/ ou /p/, représentés respectivement par les lettres 'b' ou 'p'. S'ajoutent à ce matériel 192 items distracteurs ne contenant aucun des deux phonèmes cibles.

3.1.2 Procédure expérimentale

Les sujets devaient détecter, le plus vite possible, les phonèmes /b/ et /p/.
Un même mot n'était présenté qu'une seule fois à chaque sujet, pour la détection de /p/ ou de /b/; par exemple, le mot "absurde" est présenté à la moitié des sujets pour la détection de /b/ et à l'autre moitié pour la détection de /p/.
Les temps de réponse étaient mesurés en millisecondes à partir du début du son cible (du relâchement de l'explosion).

3.1.3 Sujets

Soixante étudiants en psychologie de l'université René Descartes ont participé à cette expérience.

3.2 Résultats et discussion

3.2.1 Taux de détection

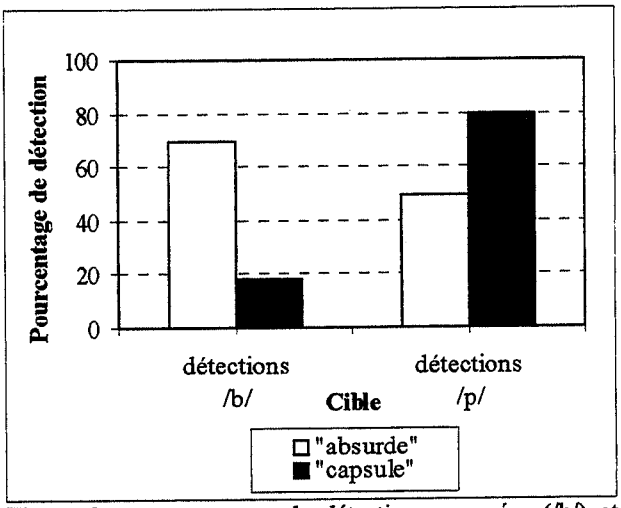


Figure 3 : pourcentages de détections erronées (/b/) et correctes (/p/) dans les mots de type "absurde" et "capsule".

Les résultats sont extrêmement nets. Les détections correctes de /p/ sont significativement plus nombreuses dans "capsule" que dans "absurde". De plus, /b/ est détecté plus souvent que /p/ dans "absurde" (le taux de détection de /p/ dans "absurde" atteint quand même 49%).
Ces résultats suggèrent que les réponses des sujets sont déterminées, dans une situation de traitement en temps réel, par les représentations phonologiques abstraites, et non de surface.

3.2.2 Temps de réponse

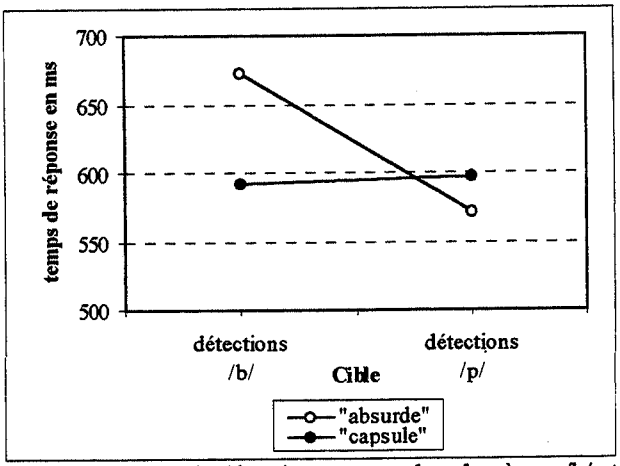


Figure 4 : temps de détections moyen des phonèmes /b/ et /p/ pour "absurde" et "capsule".

L'analyse des temps de réponse montre que les détections erronées de /b/ dans "absurde" sont significativement plus longues que les détections correctes de /p/ dans "absurde" comme dans "capsule". Ceci suggère encore

une intervention de représentations linguistiques dans le traitement de ces phonèmes.
Ceci montre que lorsqu'il n'y a pas congruence entre les deux types de représentation (de surface *vs* abtraite), le temps d'activation des représentations finalement utilisées (les représentations phonologiques abstraites) est allongé.

4. DISCUSSION GENERALE

Cette étude montre que les auditeurs s'appuient sur la forme linguistique des mots lorsqu'ils traitent des mots parlés en temps réel, et non sur leur substance sonore. En revanche, ils perçoivent bien les sons selon leur forme acoustique lorsque ceux-ci sont isolés de leur contexte lexical.
Les représentations phonologiques abstraites semblent donc interférer avec celles de surface, au point de les dominer. Quelle est la nature de cette représentation phonologique abstraite? Il pourrait s'agir de la représentation orthographique des mots, qui serait activée même en modalité auditive de présentation, comme l'ont déjà suggéré différentes études : voir par exemple [Sei79], [Taf85] [Dij95]. Il a aussi été montré qu'en modalité visuelle, les représentations phonologiques sont activées : voir par exemple [Ord87], [Jar91], [Zie97].
Mais il pourrait aussi s'agir de représentations de nature morphophonologique. Les items utilisés dans cette étude sont tous préfixés, ou du pseudo-préfixés par les préfixes {ab-}, {ob-} ou {sub-} dont la représentation phonologique abstraite est bien /ab/, /ob/, et /sub/, indépendamment de la racine qui les suit. Les items expérimentaux sont d'ailleurs morphologiquement reliés à des mots préfixés par ces mêmes préfixes, dans lesquels la lettre 'b' se prononce bien /b/ (comme "abjurer", "abdiquer"...). Par exemple, le mot "obtenir" a une représentation morphophonologique composée du préfixe {ob-} et de la racine {tenir}.
On peut donc aussi supposer que cette représentation morphophonologique est responsable des effets observés.
Quoiqu'il en soit, cette étude montre, dans une perspective plus générale, que nous ne percevons pas le monde extérieur tel qu'il est, mais que nous l'interprétons à l'aide des différentes représentations que nous avons stockées en mémoire. Si les représentations activées après le traitement d'un énoncé parlé sont bien déclenchées par la stimulation objective, elles sont loin d'en être une simple copie.

Bibliographie

[Car74] Carton, F. (1974), Introduction à la Phonétique du Français, Bordas.

[Dij95] Dijkstra, T., Fieuws, S. & Roelofs, A. (1995). "Orthographic effects on phoneme monitoring".*Canadian Journal of Experimental Psychology*, Vol. 49, pp. 264-271.

[Hal98] Hallé, P. A., Segui, J., Frauenfelder, U. & Meunier, C. (1998). " The processing of illegal consonant clusters. A case of perceptual assimilation." *Journal of Experimental Psychology : Human Perception and Performance*, Vol. 24, Perception & Psychophysics. 1-17.

[Jar94] Jared, D., & Seidenberg, M.S. (1991). "Does word identification proceed from spelling to sound to meaning?" *Journal of Experimental Psychology: General*, Vol. 120, pp. 358-394.

[Ord87] Van Orden, G. C. (1987). "A rows is a rose: Spelling, sound and reading". *Memory and Cognition*, Vol. 15, pp. 181-198

[Rig67] Rigault, A. (1967), "L'assimilation de sonorité en Français : Etude Acoustique et Perceptuelle", 6è ICPhS, Prague.

[Sei79] Seidenberg, M. S. & Tanenhaus, M. K.(1979). "Orthographic effects on rhyme monitoring". *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 5, pp. 546-554.

[Taf85] Taft, M. & Hambly, G. (1985)."The influence of orthography on phonological representations in the lexicon". *Journal of Memory and Language*, Vol. 24, pp. 320-335.

[Zie97] Ziegler, J. C., Van Orden, G. C. & Jacobs, G. C. (1997). "Phonology can help or hurt the perception of print". *Journal of Experimental Psychology : Human Perception and Performance*,Vol. 23, pp. 845-860.

Nature des représentations prosodiques en mémoire auditive

Juliette CLEMENT (1) & Claire GERARD (2)

(1) Laboratoire de Psychologie Expérimentale - URA CNRS 316 - Univ. R. Descartes, PARIS V
28, Rue Serpente - 75006 PARIS - jclement@psycho.univ-paris5.fr
(2) Laboratoire Langage et Communication - URA CNRS 1067
Département de Psychologie - 95, Avenue du Recteur Pineau - 86000 POITIERS

ABSTRACT

Similarities between the prosodic aspects of speech and certain aspects of music (in terms of surface structure) have directly inspired several models of prosodic structure, but the relationship between speech and music processing have rarely been tested empirically. Furthermore, neuropsychologic data seem to show that processing of language and music are largely independent, and support a modular organisation hypothesis of auditory memory. More specifically, we could postulate sub-modules for auditory "primitives" processing. In order to study prosodic features processing, 3 experiments have been conducted using an interference paradigm. The result argue for a separation between elementary and integrated processing of prosodic features.

1. INTRODUCTION

La prosodie, aspect exclusivement sonore et "musical" du langage, est constituée par l'évolution temporelle de trois paramètres acoustiques : durée des sons de parole et des silences, intensité, et fréquence fondamentale (F0) des sons voisés; ces mêmes paramètres se retrouvent en partie dans la musique, sous la forme du rythme et de la mélodie.

A un niveau pragmatique, la prosodie est utilisée pour marquer acoustiquement les formes illocutoires (intention communicative sous-jacente à l'énoncé, non exprimée par le contenu lexical). Ainsi, elle aurait un rôle central dans la façon dont le locuteur produit son énoncé [Cut83; Fon91; Lev89]. Dans nos recherches développementales antérieures [Clé96; Gér,révision], nous avons montré une certaine dépendance entre les paramètres prosodiques (durée, F0) dans le traitement des formes illocutoires (i.e. assertive, interrogative totale, joie, tristesse, ...). A chacune correspondait une configuration spécifique des paramètres prosodiques. Nous avons mis en évidence que les locuteurs produisaient et les auditeurs identifiaient les formes illocutoires sur la seule base des variations prosodiques, confirmant ainsi nos hypothèses sur l'existence de représentations mentales des formes prosodiques ("patterns" prosodiques). De plus, nos études ont révélé un décalage chronologique dans la maîtrise des deux paramètres prosodiques. Les enfants contrôlaient la production des paramètres intonatifs avant de contrôler réellement la production des paramètres temporels, et les configurations d'ensemble des deux paramètres prosodiques correspondaient au stade achevé du développement. Ces études nous ont amenés à réfléchir sur la nature des représentations prosodiques et nous nous sommes demandées si ces représentations étaient organisées et stockées en mémoire auditive sous la forme de configurations intégrées ou si les différents paramètres prosodiques étaient traités et stockés par des systèmes différents ?

A un niveau superficiel de ressemblance, la prosodie serait une base commune entre la parole et la musique [Cla89], les paramètres acoustiques se retrouvant dans la musique. Les similarités entre les aspects prosodiques de la parole et certains aspects de la musique (en termes de structure de surface) ont directement inspiré des modèles de la structure prosodique, mais les relations entre les traitements de la parole et de la musique ont rarement été testées empiriquement. Les données neuropsychologiques tendent à montrer que traitements du langage et de la musique sont globalement indépendants, et appuient l'hypothèse d'une organisation modulaire de la mémoire auditive, puisque les systèmes modulaires sont fonctionnellement modulaires mais aussi physiquement [Fod86]. Ainsi, Peretz & Morais ont observé des déficits dans le traitement des informations auditives (i.e. amusies sans aphasies et aphasies sans amusies)[Per89]. Des lésions (sur des patients) et des activations (mesurées avec une tomographie par émission de positons sur des sujets sains) ont été observées dans des zones cérébrales spécifiques lors de différents traitements auditifs [Zat92 ; Zat93 ; Zat94]. D'une manière générale, l'hémisphère gauche serait spécialisé dans le traitement de la parole tandis que l'hémisphère droit serait spécialisé dans le traitement de la musique, mais ces observations résultaient de recherches essentiellement faites à partir de l'écoute de mélodies. En ce qui concerne plus spécifiquement le traitement musical, Peretz pense que l'organisation des hauteurs et l'organisation temporelle sont plutôt considérées comme deux candidats distincts à la modularité à partir de l'observation du phénomène de double dissociation. Certains patients étaient encore capables de traiter la composante rythmique des séquences musicales et non la composante mélodique, et inversement pour d'autres patients [Per94]. Au niveau acoustique, nous retrouvons ces mêmes dimensions dans

l'aspect sonore du langage, la prosodie de la parole. A ce niveau de surface, nous pouvons douter d'une séparation catégorique entre les deux domaines musical et verbal, même si par exemple certains aspects de l'organisation des hauteurs sont clairement spécifiques au domaine musical (i.e. les échelles musicales et les tonalités différentes). Il est raisonnable de penser que le traitement des paramètres acoustiques n'est pas spécifique à la musique mais participe également à la perception de la prosodie de la parole [Per94]. Les organisations rythmiques seraient traitées dans l'hémisphère gauche et la hauteur tonale serait traitée dans l'hémisphère droit [Per94 ; Zat93].

Certaines données psychologiques sont en accord avec l'hypothèse d'une organisation modulaire de la mémoire auditive. A l'origine, Deutsch [Deu70] a étudié la mémoire immédiate de la hauteur tonale avec un paradigme d'interférence et elle a démontré une certaine organisation de la mémoire auditive. Les sujets avaient pour tâche de décider si deux sons purs séparés de 5 secondes étaient identiques ou différents en hauteur. Entre le son-test et le son de comparaison étaient intercalés soit 6 autres sons purs, soit 6 chiffres lus à ignorer, ou encore 6 chiffres lus mais à se rappeler. Que les sujets aient à effectuer une tâche ou non sur les chiffres, les pourcentages de réussite étaient similaires et très élevés, par contre les performances étaient dégradées dans le premier cas. La mémoire de la hauteur tonale était donc perturbée par la présence d'autres sons purs mais pas par la présence d'items verbaux. Ainsi, cet effet ne serait pas du à des facteurs généraux, comme le blocage des auto-répétitions ou la capacité limitée de stockage. Semal & Demany ont repris le même paradigme d'interférence et ont montré que le traitement de la hauteur tonale était spécifique et sourd au traitement du timbre. Ainsi, ils ont également fait l'hypothèse d'une organisation modulaire de la mémoire auditive [Sem91]. Dans un domaine plus linguistique, le traitement de la parole serait également spécifique, sourd aux autres sons, où seul le code linguistique auditif n'a pas besoin d'être recodé, avec un système direct de traitement [Sur93; Cro94]. Les observations précédentes sont à nuancer car d'autres auteurs réfutent l'idée d'une séparation tranchée dans le traitement des attributs auditifs. Serafine suppose que parole et musique sont intégrées à un degré important dans la mémoire des chansons, car elles s'aident mutuellement et il est difficile de se rappeler l'une sans l'autre [Ser84; Per89; Per94].

A partir de l'ensemble de ces données, nous nous sommes demandées à quel niveau d'intégration était traitée la prosodie par rapport aux traitements de la parole et de la musique dans la mémoire auditive, et plus spécifiquement quelles étaient les relations entre les différentes composantes prosodiques. A des niveaux de traitement précoces, les paramètres acoustiques pourraient avoir un statut de "primitives" auditives, c'est à dire des composants élémentaires à partir desquels sont construites les représentations auditives. Dans cette optique, nous adoptons l'hypothèse générale d'une organisation modulaire de la mémoire auditive et nous supposons l'existence de sous-modules attachés au traitement des différentes dimensions acoustiques.

Nous avons alors étudié le traitement perceptif prosodique de sons de parole, à différents niveaux (phrastique et lexical). Pour montrer l'existence de traitements spécifiques des traits prosodiques (durée, F0), nous avons donc adapté le paradigme d'interférence (AXB) de Deutsch [Deu70]. Le principe reste le même mais la tâche de comparaison porte sur des séquences verbales test (A) et des séquences verbales de comparaison (B). Les séquences interférentes (X) durent toujours 5 secondes mais elles sont de nature différente d'une expérience à l'autre (cf. ci-dessous). Nous analysons les temps de réponses ainsi que les pourcentages de bonnes réponses. Les comparaisons effectuées par les sujets portaient tantôt sur des variations intonatives, tantôt sur des variations temporelles.

2. ROLE DES INTERFERENCES VERBALES ET MUSICALES DANS LE TRAITEMENT DE LA PROSODIE

2.1. Méthode

Niveau phrastique

Le matériel expérimental était composé d'une phrase courte et simple "il fait froid" produite de façon différente par un locuteur humain. Les variations portaient sur un seul paramètre à la fois, l'autre étant maintenu constant ("moyen") : pour l'intonation, le locuteur avait produit l'énoncé sur des tons grave (F0 moyenne de 190Hz), moyen (230Hz) et aigu (270Hz); pour la durée, les débits étaient rapide (durée totale de 800ms), moyen (1000ms) et lent (1230ms). Chaque variation prosodique était environ de +/- 20 % par rapport à l'énoncé "moyen" (230Hz et 1s). Les séquences interférentes étaient composées de phrases de parole produites par le même locuteur (P : e.g. "Ce matin, quand je me suis levée, il s'est mis à pleuvoir"), de phrases musicales (M : petites mélodies comprenant entre 15 et 20 notes), ou de silence (S). Cette dernière condition sert de condition contrôle pour comparer les résultats obtenus.

Niveau lexical

Le matériel expérimental était composé de mots bisyllabiques fréquents dans la langue française (plus de 100 occurrences par million). Ils étaient de structure CVCV, sans e muet, sans double consonne, ... Enoncés d'abord sur un ton et un débit "moyen" (F0 moyenne de 225Hz et durée totale d'environ 600ms) par un locuteur humain, ils ont été transformés artificiellement par logiciel informatique. Pour la F0, le contour mélodique était translaté de +/- 10% par rapport au ton moyen; pour le débit d'élocution, les mots étaient condensés ou étirés dans le temps de +/- 12.5% par rapport au débit moyen. Les séquences interférentes (X) étaient également composées soit de silence (S), soit de parole (P) ou encore

de musique (M) et elles étaient de complexité différente pour les deux dernières. L'interférence P pouvait être une séquence de monosyllabes, une séquence de bisyllabes, ou une phrase, l'interférence M pouvait être une suite de notes aléatoires, une séquence de binotes à contours plat, montant, et descendant, ou une petite mélodie.

2.2. *Résultats*

Les tableaux 1 et 2 ci-dessous présentent les différents résultats observés lors des tâches de comparaison prosodique au niveau de la phrase (table 1) et au niveau du mot (table 2). D'une manière générale, les pourcentages de bonnes réponses sont plus importants systématiquement lors des comparaisons intonatives que lors des comparaisons temporelles. Donc, malgré le souci de rendre plus difficiles les comparaisons sur la F0 en utilisant une gamme de variation intonative plus restreinte pour l'expérience sur les mots, les jugements temporels restent plus difficiles. L'analyse des temps de réponses va parfois dans le même sens : les sujets mettent plus de temps pour répondre à une comparaison temporelle que pour répondre à une comparaison intonative lors de la tâche des jugements prosodiques sur les mots

Table 1 : Temps de réponse et pourcentages de bonnes réponses à la tâche de comparaison prosodique sur des phrases en fonction des paramètres acoustiques (F0 et Débit) et des séquences interférentes (Silence, Parole ou Musique).

Phrases	F0				Dé			
	S	P	M	Moy.	S	P	M	Moy.
TR (ms)	617	691	909	739	549	644	750	648
BR (%)	100	90,6	81,3	90,6	100	71,9	78,1	83,3

Moyennes	S	P	M
TR (ms)	583	668	830
BR (%)	100	81,3	79,7

La nature des séquences interférentes entraîne également un effet global sur les réponses des sujets. Les pourcentages de bonnes réponses sont systématiquement différents, la condition silence donnant toujours lieu au plus fort score de bonnes réponses. L'analyse des temps de réponse va dans le même sens. Mais quelle que soit l'expérience, les pourcentages de bonnes réponses ne sont pas vraiment différents en fonction de la nature verbale ou musicale de l'interférence (pour les interférences P versus M, les pourcentages sont respectivement de 81.3% et de 79.7% dans le tableau 1, ils sont de 74% et de 74.9% dans le tableau 2). En outre, les pourcentages de bonnes réponses pour les comparaisons intonatives ou temporelles ne sont pas exactement répartis de la même façon dans les deux expériences.

Pour l'expérience sur les phrases, les comparaisons intonatives sont plus dégradées par l'interférence musicale et les comparaisons temporelles sont plus dégradées par l'interférence verbale. De plus, d'une manière générale, l'interférence musicale a majoritairement augmenté les temps de réponses. Pour l'expérience sur les mots, par contre, les comparaisons entre les interférences verbales et musicales ne sont pas significatives. Les sujets, cette fois-ci, n'ont pas été plus gênés par la parole ou par la musique lors des deux types de comparaisons prosodiques différentes.

Table 2 : Temps de réponse et pourcentages de bonnes réponses à la tâche de comparaison prosodique sur des bisyllabes en fonction des paramètres acoustiques (F0 et Débit) et des séquences interférentes (Silence, Parole ou Musique).

Mots	F0				Dé			
	S	P	M	Moy.	S	P	M	Moy.
TR (ms)	486	549	522	519	612	597	592	600
BR (%)	90	84,7	88,9	87,9	69,7	63,3	60,8	64,6

Moyennes	S	P	M
TR (ms)	549	573	557
BR (%)	79,9	74	74,9

2.3. *Discussion*

Les jugements temporels sont toujours plus difficiles que les jugements intonatifs. Ces résultats systématiques vont dans le même sens que le décalage chronologique de la maîtrise des paramètres prosodiques observés dans nos études sur le développement des connaissances prosodiques chez l'enfant. La dimension temporelle reste un obstacle difficile à surmonter.

A un niveau de traitement supérieur (la phrase), nous avons pu mettre en évidence un rapport entre l'interférence musicale et les traitements intonatifs ainsi qu'entre l'interférence verbale et les traitements temporels. Ce genre de résultats est grossièrement en accord avec les différentes spécialisations hémisphériques présentées antérieurement. Mais si nous passons à un niveau plus élémentaire de traitement (niveau lexical), nous nous apercevons qu'il n'y a plus d'interférence spécifique. Les différents jugements prosodiques ne se situeraient donc pas au même niveau de traitement que l'opposition parole versus musique. Et il faudrait peut-être chercher à un niveau plus élémentaire (de l'ordre des "primitives") pour montrer une organisation modulaire concernant les traitements spécifiques effectués sur les paramètres prosodiques.

3. TRAITEMENTS SPECIFIQUES DES PARAMETRES PROSODIQUES ?

3.1. *Méthode*

Le matériel expérimental est composé de la même liste de mots bisyllabiques utilisée lors de l'expérience précédente. A l'exception près que nous avons encore essayé d'affiner les différences entre les paramètres prosodiques par l'intermédiaire d'une étude sur les seuils différentiels de perception des différents paramètres (non présentée ici).

Pour le paramètre intonatif, le contour mélodique est translaté de +/- 6% par rapport au ton moyen (environ 225Hz); pour le débit d'élocution, les mots ont été condensés ou étirés dans le temps de +/- 15% par rapport à la durée totale des mots (600ms). Cette fois-ci les séquences interférentes (X) sont composées de silence ou d'une suite de mots bisyllabiques. Ces mots varient également sur un paramètre à la fois, ils varient donc dans une gamme de 12% à partir d'une F0 moyenne de 225HZ; et dans une étendue temporelle de 30% à partir d'une durée moyenne de 600ms.

Il nous a semblé intéressant de tester spécifiquement s'il existait une interférence intonative dans les comparaisons de hauteur et inversement s'il existait une interférence temporelle lors de jugements temporels. Nous avons utilisé le même support matériel (mots bisyllabiques) pour que les interférences et les séquences à comparer soient au même niveau de traitement. Si l'hypothèse d'une organisation modulaire est valable et a fortiori celle de l'existence de sous-modules spécifiques au traitement des différents attributs acoustiques, alors nous nous attendons à obtenir une interaction croisée entre la nature des séquences interférentes et les différentes variations prosodiques à comparer. En d'autres termes, nous pensons que le traitement des informations temporelles interférentes dégradera moins les performances lors de comparaisons intonatives et inversement le système utilisé pour les comparaisons temporelles sera moins gêné par le traitement des variations de hauteur. Ce genre de résultats serait en accord avec l'idée de "primitives" auditives à un niveau précoce de traitement, indépendamment des traitements de parole ou de musique. Et il ne serait pas en désaccord avec l'idée que ces différents attributs sonores peuvent être intégrés en configurations d'ensemble à des niveaux supérieurs de traitement. Ces niveaux supérieurs de traitement peuvent correspondre à un registre proprement linguistique réunissant les connaissances nécessaires à la production et à la perception des formes illocutoires par exemple ou ils peuvent correspondre à un registre proprement musical réunissant des connaissances sur les formes mélodiques et les organisations rythmiques.

BIBLIOGRAPHIE

[Cla89] Clarke, E.F. (1989). "Considérations sur le langage et la musique", in S. McAdams & I. Deliège (Eds.), La musique et les Sciences Cognitives, Bruxellles: Mardaga, pp. 23-42.

[Clé96] Clément, J. & Gérard, C. (1996). "Programmation de la production et anticipation de la perception des formes prosodiques", Actes des 21èmes Journées d'Etude sur la Parole, Avignon, pp. 199-202.

[CRO94] Crowder, R.G. (1994). "La mémoire auditive", in S. McAdams & E. Bigand (Eds.), Penser les sons, Paris: PUF.

[Cut83] Cutler, A. & Ladd, D.R. (1983). Prosody : Models and measurements. Springer-Verlag, Berlin, Heidelberg, pp. 1-9.

[Deu70] Deutsch, D. (1970). "Tones and numbers: Specificity of interference in immediate memory", Science, Vol. 168, pp. 1604-1605.

[Fod86] Fodor, J.A. (1986). La modularité de l'esprit, Essai sur la psychologie des facultés. Les éditions de minuit.

[Fon91] Fonagy, I. (1991). La vive voix, essais de psycho-phonétique. Bibliothèque scientifique Payot.

[Gér] Gérard, C. & Clément, J. "The structure and development of french prosodic representations" (en cours de révision pour la revue Language and Speech).

[Kru92] Krumhansl, C.L. & P. Iverson (1992). "Perceptual interactions between musical pitch and timbre", Journal of experimental psychology : human, perception and performance, Vol. 18 (3), pp. 739-751.

[Lev89] Levelt, W.J.M. (1989). Speaking. From intention to articulation. Cambridge, Massachusetts: The MIT Press.

[Per94] Peretz, I. (1994). "Les agnosies auditives: une analyse fonctionnelle", in S. McAdams & E. Bigand (Eds.), Penser les sons, Paris: PUF, pp. 215-248.

[Per89] Peretz, I. & Morais, J. (1989). "La musique et la modularité", in S. McAdams & I. Deliège (Eds.), La musique et les Sciences Cognitives, Bruxellles: Mardaga, pp. 393-415.

[Sem91] Semal, C. & Demany, L. (1991). "Dissociation of pitch from timbre in auditory short-term memory", Journal of the Acoustical Society of America, Vol. 89, pp. 2404-2410.

[Ser84] Serafine , M.L., R.G. Crowder & B. Repp (1984). "Integration of melody and text in memory for songs", Cognition, Vol. 16, pp. 285-303.

[Sur93] Surprenant, A.M., M.A. Pitt & R.G. Crowder (1993). "Auditory recency in immediate memory", The quarterly journal of experimental psychology, Vol. 46A (2), pp. 193-223.

[Zat92] Zatorre, R.J., Evans, A.C. & Meyer, E. & Gjedde, A. (1992). "Lateralization of phonetic and pitch discrimination in speech processing", Science, vol. 256, pp. 846-849.

[Zat93] Zatorre, R.J. (1993). "Fonctions cérébrales et traitements auditifs complexes". Revue de Neuro-psychologie, Vol. 3 (n°4), pp. 407-425.

[Zat94] Zatorre, R.J., Evans, A.C. & Meyer, E. (1994). "Neural mechanisms underlying melodic perception and memory for pitch", Journal of Neuroscience, Vol. 14 (n°4), pp. 1908-1919.

Groupes consonantiques illégaux et segmentation de la parole.

Olivier Crouzet & Nicole Bacri

Laboratoire de Psychologie Expérimentale
Université René Descartes - CNRS URA 316
28, rue Serpente, 75006 Paris
Tél.: 01 40 51 98 65 - Fax: 01 40 51 70 85
e-mail: olivier.crouzet@psycho.univ-paris5..fr

ABSTRACT

It has been argued that phonological constraints could play a role in speech segmentation [Bre96][Chu87]. A previous experiment [Cro97] led us to claim that constraints on legal and illegal consonant clusters were used by the system to segment the speech signal. However, many different models of speech perception could predict this kind of results without invoking the use of phonological constraints. Our results allow us to reject some of these explanations. The ways these constraints could be implemented into the speech processing system are reviewed.

1. INTRODUCTION

Le signal de parole est une onde acoustique relativement exempte de marques de frontières entre les mots. Nous parvenons cependant sans aucune difficulté à isoler les divers éléments de la chaîne parlée. L'un des processus qui semble permettre d'aboutir à cette segmentation est celui d'activation interactive tel qu'il est mis en œuvre dans le modèle TRACE [McC86]. Selon ces auteurs, le système met en œuvre des détecteurs de traits qui vont activer des phonèmes, ceux-ci activant à leur tour les mots du lexique. Ce sont les processus de compétition entre candidats lexicaux qui conduisent à l'identification des mots et, de fait, à l'extraction des frontières lexicales. Ainsi, la segmentation du signal est une simple *conséquence* de l'accès au lexique.

Cependant, d'autres processus peuvent être envisagés qui ne remettent pas nécessairement en cause l'idée de compétitions. Ainsi, Norris, McQueen & Cutler [Nor95] proposent d'implémenter des procédures de segmentation prélexicale fondées sur une information métrique dans un modèle mettant en œuvre des compétitions entre les candidats lexicaux (Shortlist, [Nor94]). Ces procédures sont fondées sur l'accentuation. On peut également supposer, comme le fait Church [Chu87], que les contraintes phonologiques segmentales -qui portent sur les séquences de phonèmes- puissent intervenir dans les processus de traitement de la parole. En effet, chaque langue obéit à des "lois" sur l'organisation séquentielle des sons qu'elle utilise. Ces contraintes pourraient servir de guide au système afin de faire des hypothèses quant aux points du signal qui pourraient constituer une frontière lexicale.

Pour notre part, nous nous sommes intéressés aux contraintes portant sur les groupes de consonnes. Dans chaque langue, il existe des groupes consonantiques (ou *clusters*) qui peuvent apparaître en début ou en fin de mot. Ces groupes sont dits légaux [Del95]. Ceux qui ne peuvent pas apparaître dans cette position sont définis comme illégaux dans cette langue. Si ces contraintes phonologiques peuvent être utilisées par le système pour repérer des frontières de mots, la comparaison de ces deux types de groupes consonantiques dans une tâche faisant intervenir les processus de segmentation de la parole devrait mettre en évidence une différence dans les performances des sujets.

Dans une précédente expérience [Cro97], nous avions utilisé la tâche d'*extraction de mot* (Word Spotting, [Cut88]) dans laquelle les sujets doivent détecter des mots monosyllabiques au début d'une séquence de deux syllabes sans signification (ils peuvent par exemple entendre la séquence /bagryʒ/ et doivent dans ce cas détecter le mot "bague"). Afin d'étudier le rôle des clusters illégaux dans la segmentation de la parole, les mots étaient suivis d'une consonne qui pouvait donner lieu à un groupe consonantique légal (p.ex. /bag/ était suivi d'une syllabe commençant par un /r/) ou illégal (/bag/ était alors suivi d'une syllabe débutant par un /v/ ou un /d/). La détection des mots était plus rapide pour les groupes illégaux que pour les groupes légaux. Ainsi, "bague" était détecté plus rapidement dans /bagvyʒ/ que dans /bagryʒ/. Nous avons interprété cet effet comme le reflet de l'utilisation de contraintes phonologiques par le système de traitement. Le système aurait accès aux contraintes portant sur les séquences de consonnes légales dans la langue et postulerait une frontière lexicale quant il rencontrerait une séquence qui ne respecte pas ces contraintes. Cependant, un certain nombre de variables n'avaient pas été prises en compte dans la constitution du matériel et d'autres modèles peuvent rendre compte de cet effet.

2. MODELES THEORIQUES

Le modèle COHORT [Mar78] est le premier à introduire l'idée de compétitions lexicales. Lorsqu'on traite un signal de parole, les 150 premières millisecondes

conduiraient à l'activation des membres de la "cohorte". Par la suite, le traitement progressif de l'information permet de supprimer de cette cohorte les candidats déviants ; l'ultime candidat correspond alors au mot qui est identifié. Ainsi, à mesure que l'information s'accumule, un certain nombre de candidats sont supprimés de la cohorte. Il est possible que, dans le matériel que nous avions utilisé, au point où la consonne qui fait que le cluster est légal ou non -la consonne critique- apparaît, les séquences légales eussent donné lieu à un nombre de compétiteurs plus grand que les séquences illégales. Notre matériel n'ayant pas été contrôlé sur cet aspect, il est possible que les mots qui apparaissaient dans une séquence comportant un groupe illégal aient été détectés plus rapidement en raison d'un nombre plus faible de compétiteurs.

Le modèle TRACE [McC86] reprend l'idée de compétitions lexicales en introduisant une procédure d'alignement exhaustif. COHORT engendrait une cohorte de candidats sur le début d'un mot. Il fallait avoir identifié un mot pour être en mesure de générer une cohorte pour le mot suivant. Ceci posait un certain nombre de problèmes, notamment pour les mots courts, qui sont résolus avec TRACE. L'alignement exhaustif consiste à générer un "réseau" d'hypothèses lexicales pour chaque segment du signal de parole. Par exemple, la séquence /bato/ va conduire à l'activation de "Bijou" mais également de "vAllée", "piTon", "pEAU", etc. C'est le jeu des activations et inhibitions entre les diverses unités du réseau qui conduit à l'identification des mots présents dans l'énoncé. L'une des prédictions que l'on peut dériver de TRACE est que si un groupe consonantique n'existe pas dans le lexique, seuls des mots ne chevauchant pas ce cluster pourront être activés ! La séquence légale va donc nécessairement permettre l'activation d'un plus grand nombre de candidats et sera d'autant plus difficile à segmenter.

Plus récemment, certaines propositions ont été avancées en faveur de l'utilisation des contraintes distributionnelles de la langue dans la segmentation de la parole [Bre96][Saf96]. On peut ici envisager que la fréquence d'occurrence des groupes de consonnes dans le lexique joue un rôle dans la segmentation. Un groupe apparaissant fréquemment pourrait être plus difficile à segmenter qu'un groupe rare en raison de sa plus grande probabilité d'apparition à l'intérieur d'un mot.

Si l'on veut étudier l'utilisation des contraintes phonologiques dans la segmentation de la parole, il devient dès lors nécessaire de contrôler un nombre important de variables. Dans la préparation de cette expérience, nous avons poursuivi trois objectifs :

(1) supprimer les biais potentiellement engendrés par les phénomènes de compétitions lexicales.

(2) utiliser uniquement des groupes de consonnes qui existent dans le lexique français.

(3) contrôler la fréquence d'occurrence des groupes de consonnes utilisés.

3. METHODE

L'objet de cette expérience est d'étudier le rôle des groupes consonantiques illégaux dans la segmentation de la parole. Un certain nombre de contrôles ont été réalisés afin de préciser dans quelle mesure les locuteurs sont capables d'utiliser les contraintes de leur langue pour segmenter le signal de parole.

3.1 Sujets

Vingt-quatre étudiants de second cycle à l'Université René Descartes - Paris V ont participé à l'expérience. Ils étaient tous de langue maternelle française et ne souffraient d'aucun trouble auditif connu.

3.2 Matériel

34 mots monosyllabiques ont été prononcés au début de séquences de deux syllabes sans signification. La consonne finale du mot était toujours une obstruante. La consonne suivante pouvait donner lieu à trois catégories de groupes consonantiques (obstruante-liquide comme /kl/, obstruante-fricative comme /ks/ et obstruante-obstruante comme /kt/).

Tableau 1 : Exemples de stimuli utilisés dans les trois conditions expérimentales (obli : obstruante-liquide, obfr : obstruante-fricative, obob : obstruante-obstruante).

obli	obfr	obob
sac luze	sac chuze	sac puze
/saklyz/	/sakʃyz/	/sakpyz/

Nous avons calculé, à partir des tables de BRULEX, un indice de probabilité d'apparition dans le lexique en utilisant une mesure de fréquence des clusters pondérée par la fréquence des mots dans lesquels ils apparaissent. Il ressort de cette analyse statistique que les trois types de groupes consonantiques utilisés présentent en moyenne des probabilités d'occurrence différentes. En raison des caractéristiques distributionnelles du lexique français, il n'a pas été possible de maintenir constante la fréquence des groupes de consonnes. Nous verrons cependant dans la discussion qu'il a été possible d'estimer l'impact de ce facteur sur les processus de segmentation du signal de parole.

Par ailleurs, nous n'avons utilisé que des groupes de consonnes existant dans les mots du lexique. Ceci permet de rejeter l'idée -dérivée de TRACE [McC86]- que cet effet, s'il se maintient, est lié à la non-existence des clusters dans le lexique. Enfin, les stimuli utilisés ont été construits dans le but de contrôler les phénomènes de compétition prédits par COHORT [Mar78]. Pour chaque séquence auditive, au point auquel apparaît la consonne critique, il n'y a plus de compétiteur possible dans le lexique. Ceci est vrai pour les trois types de clusters

utilisés. Par exemple, pour les trois exemples donnés dans le tableau 1, il n'existe pas de mot français commençant par /sakl/, par /sakʃ/ ou par /sakp/. Ce matériel nous permet donc de contraindre les interprétations possibles de l'effet observé dans [Cro97].

Trente-quatre items ne contenant aucun mot ont également été utilisés (p.ex. /driglun/). Les stimuli ont été enregistrés sur DAT puis digitalisés sur 16 bits à une fréquence d'échantillonnage de 16 kHz. La fréquence fondamentale était maintenue relativement constante à l'enregistrement.

3.3 Procédure Expérimentale

Nous avons utilisé la tâche d'extraction de mot [Cut88] dans laquelle les sujets doivent détecter des mots en position initiale de non-mot. Afin que chaque sujet soit soumis aux trois types de groupes de consonnes tout en n'entendant un même mot qu'une seule fois, trois listes ont été créées. Chaque sujet était soumis aléatoirement à l'une des trois listes expérimentales. Les sujets étaient prévenus qu'ils allaient entendre des non-mots et que certains de ces non-mots commenceraient par un mot. Leur tâche consistait à appuyer le plus rapidement possible sur un bouton réponse dès qu'ils détectaient un mot inséré dans un non-mot. Après avoir appuyé, ils devaient dire quel mot ils avaient identifié. Les stimuli étaient présentés à un niveau d'écoute confortable (environ 65dB). Les temps de réaction étaient mesurés depuis l'occlusion de la consonne finale du mot. L'expérience commençait par une phase de familiarisation avec la tâche.

4. RESULTATS

Les données de 23 sujets ont été analysées (F_1 par sujet, F_2 par item). Les temps de réaction supérieurs à 1400 ms et les omissions (15% des données) n'ont pas été intégrés dans l'analyse. Le facteur liste n'interagissant pas avec le type de groupe consonantique, les analyses présentées ne le prennent pas en compte.

Les temps d'extraction de mot varient en fonction du type de groupe de consonnes (cf. Figure 1). Les mots sont détectés plus lentement lorsqu'ils sont suivis d'une consonne donnant lieu à un groupe consonantique obstruante-liquide que lorsque cette consonne constitue un groupe obstruante-fricative (735ms vs. 706ms). Ceux-ci présentent, pour leur part, des temps de détection plus longs que les groupes obstruante-obstruante (706ms vs. 688ms). L'effet global du type de cluster est significatif par sujet ($F_1(2,44)=3.611$, $p<.05$) mais pas par item ($F_2<1$). Afin de mieux comprendre la distribution des temps de réaction, nous avons réalisé des comparaisons partielles entre les divers types de groupes consonantiques. Les temps de réaction ne diffèrent pas significativement pour la comparaison obob vs. obfr ($F_1(1,22)=1.382$, $p>.1$; $F_2<1$). Par contre, les temps de détection dans la condition obli sont significativement plus longs que dans les deux autres conditions ($F_1(1,22)=5.186$, $p<.05$; $F_2<1$).

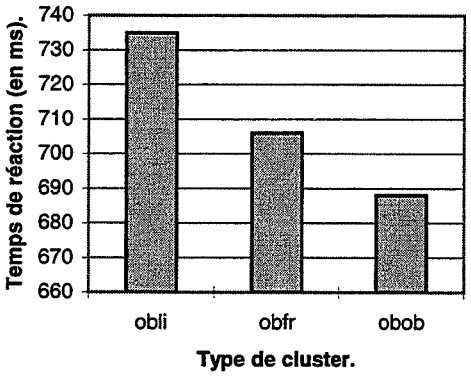


Figure 1: Temps de détection des mots (en ms) en fonction de la consonne qui les suit.

5. DISCUSSION

Notre but était d'appréhender la possibilité qu'ont les locuteurs d'utiliser les contraintes phonologiques de leur langue dans la segmentation des signaux de parole. Les résultats obtenus répliquent ceux de [Cro97]. Les locuteurs sont donc en mesure de mettre en œuvre des routines de segmentation ayant pour fondement les contraintes inhérentes à leur langue. Cette expérience permet cependant d'aller plus loin dans la compréhension de ces routines.

Tout d'abord, il apparaît clairement que l'effet obtenu dans [Cro97] n'est pas lié à des phénomènes de compétitions entre candidats lexicaux. Ainsi, quel que soit le contexte phonémique dans lequel les mots étaient prononcés, aucun compétiteur ne pouvait être activé au niveau de la consonne critique. De plus, tous les groupes utilisés dans cette expérience se retrouvent dans des mots du lexique. Ce n'est donc pas non plus l'existence ou l'inexistence de ces clusters qui contribue à l'observation de cet effet.

Le problème de la fréquence d'apparition des groupes de consonnes dans le lexique est plus délicat à résoudre. Nous n'avons pas pu contrôler ce facteur en le maintenant constant. Cependant, on n'observe pas de différence significative entre les clusters obstruante-fricative et obstruante-obstruante. Or l'analyse statistique réalisée permet, en ce qui concerne la fréquence, de situer les obstruante-fricative à une position médiane aux deux autres types de groupes. Si c'est bien leur fréquence d'occurrence qui joue, les clusters obstruante-fricative auraient dû donner lieu à des temps de détection de mot différents de ceux observés pour les clusters obstruante-obstruante. Or ce n'est pas le cas. Les temps de détection observés pour ces deux conditions ne diffèrent pas significativement. Il nous semble donc peu probable que ces effets soient liés aux caractéristiques distributionnelles de la langue.

Ces résultats confirment l'idée que le système de traitement de la parole utilise des contraintes phonologiques pour produire des attentes sur les points du signal qui pourraient constituer une frontière de mot. Cependant, même si l'on admet que ces processus sont indépendants des compétitions, qu'il sont prélexicaux et qu'ils ne font pas intervenir les caractéristiques distributionnelles de la langue, plusieurs conceptions peuvent être proposées en ce qui concerne l'implémentation de ces contraintes dans le système de traitement de la parole.

L'un des problèmes de l'étude des contraintes sur les groupes de consonnes est de choisir une définition fonctionnelle de ce qu'est un cluster illégal. En effet, une définition possible serait de dire qu'un groupe est légal s'il existe en début ou en fin de mot dans la langue [Del95]. Or un certain nombre de clusters apparaissent en Français dans ces positions et pourraient a priori être plutôt désignés comme illégaux en ce qui concerne les processus perceptifs. C'est le cas de /kt/ dans "chtonien" ou de /pt/ dans "ptérodactyle". Dell, bien qu'il les considère comme légaux sur le plan de la phonologie, remarque que ce sont néanmoins des groupes "déviants". Il est donc important de bien distinguer les contraintes de la langue et celles qui sont utilisées par le système de perception de la parole ou dans les intuitions des locuteurs. Evidemment, "chtonien" et "ptérodactyle" sont des mots extrêmement rares. Mais si l'on penche pour une définition en termes de fréquence d'apparition des clusters dans la langue et que l'on propose un seuil au-delà duquel un groupe est dit légal (ce qui pourrait alors expliquer l'absence de différence entre les obfr et les obob), le problème ne se résout pas. Par exemple, /kl/, /ks/ et /kt/ ont une probabilité d'occurrence identique. Comment définir alors ce qui pour un locuteur constituerait un groupe illégal *en tant que jouant un rôle dans les routines de segmentation* ? Il nous semble plus adéquat d'adopter une définition phonétique de ce qu'est un cluster légal en postulant que seuls les groupes obstruante-liquide, pour des raisons certainement acoustico-articulatoires [Byr96], sont des groupes légaux pour le système perceptif.

Une fois adoptée une définition formelle de la légalité des groupes de consonnes, il reste à comprendre à quel niveau de traitement ces contraintes peuvent agir. Une conception, fortement dérivée des descriptions phonologiques, consiste à envisager que le système encode une chaîne de phonèmes sur laquelle ces contraintes vont intervenir de manière descendante en indiquant à quel point de la chaîne phonémique insérer une frontière. Mais l'on peut également supposer, à partir d'une conception acoustico-articulatoire de la légalité, que ces contraintes agissent de manière ascendante avant même l'encodage d'une chaîne phonémique. Ceci irait dans le sens du modèle proposé par Marslen-Wilson & Warren (LAFF, [Mar94]) dans lequel l'accès au lexique se fait directement à partir des traits acoustico-articulatoires sans passer par un niveau de représentation phonémique. Les contraintes phonologiques interviendraient alors sur un niveau de représentation beaucoup plus précoce que celui des phonèmes et modifieraient directement la représentation même de la chaîne parlée. Ce type de conception n'est, à notre connaissance, pas envisagé dans les systèmes de reconnaissance automatique de la parole qui font en général intervenir une étape de décodage acoustico-phonétique avant l'identification des mots.

BIBLIOGRAPHIE

[Bre96] Brent M.R. & Cartwright T.A. (1996), "Distributional regularity and phonotactic constraints are useful for segmentation." Cognition, Vol.61, pp. 93-125.

[Byr96] Byrd D. (1996), "Influences on articulatory timing in consonant sequences." Journal of Phonetics, 24, 209-244.

[Chu87] Church K.W. (1987), "Phonological parsing and lexical retrieval." Cognition, Vol 25, pp. 53-69.

[Cro97] Crouzet O., Banel M-H. & Bacri N. (1997), "Intégration des informations phonotactiques et métriques dans la perception de la parole." Actes du 4ème Congrès Français d'Acoustique, pp. 345-348.

[Cut88] Cutler A. & Norris D. (1988), "The role of strong syllables in segmentation for lexical access." JEP:HPP, Vol. 14, pp. 113-121.

[Del95] Dell F. (1995), "Consonant clusters and phonological syllables in French." Lingua, Vol. 95, pp. 5-26.

[Mar78] Marslen-Wilson W.D. & Welsh A. (1978), "Processing interactions and lexical access during word recognition in continuous speech." Cognitive Psychology, Vol. 10, pp. 29-63.

[Mar94] Marslen-Wilson W.D. & Warren P. (1994), "Levels of perceptual representation and process in lexical access: Words, phonemes and features." Psychological Review, Vol. 101, pp. 653-675.

[McC86] Mc Clelland J.L. & Elman J.L. (1986), "The TRACE model of speech perception." Cognitive Psychology, Vol. 18, pp. 1-86.

[Saf96] Saffran J.R., Newport E.L. & Aslin R.N. (1996), "Word segmentation: The role of distributional cues." JML, Vol. 35, pp. 606-621.

Unité minimale de perception de la parole chez le nouveau-né : Capacité de discrimination syllabique ou moraïque ?

Hidéko YAMASAKI

Institut de Linguistique et Phonétique Générales et Appliquées - UA1027
Université de Paris III
19 rue des Bernardins - Paris (France).
E-mail : yamasaki@msh-paris.fr

Abstract : Many studies have examined speech perception in newborns, particularly the rhythmical unit of speech representation. The results of some studies indicate that newborns are able to discriminate bi-syllabic words from tri-syllabic ones.
The present experiments explore the new-borns' ability in discriminating lists of Japanese words. We know that the Japanese language possesses a rhythmic property which is not marked by the syllable but by a different unit called the "mora".
By using the High Amplitude Sucking Procedure, we observed the discrimination of this sub-syllabic unit in newborns.
The results obtained suggest that even if with reducing phonetic and prosodic variability of stimuli, newborns are not able to discriminate Japanese words distinguished by the number of moraic units. These results support the opinion that speech representation in newborns is based on the syllabic unit and that, later in development, infants acquire the unit of speech representation in his mother tongue.

Key-words : newborn, perception, rhythmical unit, syllable, mora

1. INTRODUCTION

Les recherches sur les capacités perceptives et cognitives que possède le très jeune enfant se sont développées depuis quelques dizaines d'années. Le nouveau-né n'est plus, aujourd'hui, considéré comme une page blanche, "tabula rasa". Grâce à de nombreuses études, nous savons aujourd'hui que le nouveau-né est sensible, dès ses premières heures, à la voix de sa mère [Dec80], à sa langue maternelle [Meh86] et à la rythmicité de la langue [Bij93; Ber95]. Il a été aussi montré que le nouveau-né est sensible à des indices existant à la frontière des mots, c'est-à-dire à la différence acoustique d'une séquence [mati] dans "mathématiques" et celle dans "pyjama tissé" [Chr93; Chr94].

Les travaux antérieurs ont ainsi montré que nous disposons, dès la naissance, de certaines capacités linguistiques. Afin de mieux connaître le système de traitement de la parole chez le nouveau-né, nous nous sommes intéressée à comprendre le processus de segmentation qu'il utilise dans le traitement du flux sonore, en particulier, à connaître l'unité rythmique de représentation de la parole utilisée chez le nouveau-né.

Chez les adultes (l'état stable), il a été montré que le traitement de l'unité perceptive de la parole dépendrait des propriétés rythmiques de leur langue maternelle [Meh81; Cut86; Ota93]. Quant à la perception rythmique chez le nouveau-né (l'état initial), certaines études antérieures ont montré leur capacité à discriminer les mots français bi-syllabiques de ceux tri-syllabiques [Bij93]. Les mêmes résultats ont été obtenus avec des mots japonais [Ber95]. Ces résultats ont donc confirmé leur capacité de discrimination des mots sur la base d'une **syllabe**.

Toutefois, il est à préciser que la rythmicité est en japonais, marquée non par la syllabe mais par une autre unité infra-syllabique, appelée la "more". Par exemple, dans la poésie japonaise, on compte le nombre de mores et non de syllabes. La more correspond, en japonais, dans la plupart des cas à une syllabe. Quelques unités moraïques sont sub-syllabiques : l'allongement de voyelle **/:/**, la consonne nasale **/N/**, la première partie de la consonne géminée **/Q/**:

Ex.	1) [koto]	"l'ancienne capitale"	**2** mores
	2) [ko:to]	"un manteau"	**3** mores
	[koNto]	"une comptine"	
	[kaQta]	"avoir gagné"	

Cette unité moraïque porte également l' "accent mélodique" (pitch accent). Chaque more est prononcée avec une hauteur haute (H) ou basse (B). La position de l' "accent mélodique" peut modifier la signification des mots :

Ex. 1) ['koto], Haut-Bas "l'ancienne capitale"
2) [ko'to], Bas-Haut "une chose"

Ainsi, le japonais possède une unité de segmentation qui le différencie du français (la syllabe). Le traitement des mots sur la base d'une more chez les locuteurs japonais a été confirmé dans l'étude d'Otake et al. [Ota93]. Afin d'explorer la perception de cette propriété rythmique chez le nouveau-né, plusieurs expériences ont été menées [Ber95].
Les stimuli utilisés dans ces expériences étaient des mots japonais constitués de **différents nombres** de mores, avec des phonèmes et des contours **variés**. Les résultats de cette recherche ont indiqué que les nouveau-nés français sont incapables de discriminer des mots japonais sur la base d'une "more". Il est possible que ces résultats soient dus à la **complexité** de stimuli.

Nous référant toujours à ce caractère rythmique, l'objectif de notre série d'expériences est d'explorer les capacités des nouveau-nés français à discriminer des mots japonais

sur la base d'une more dans un contexte "simplifié", à savoir la variabilité des stimuli sur le plan phonémique et prosodique est extrêmement réduite par rapport aux précédentes expériences.

2. EXPERIENCE 1

Cette expérience a pour but d'observer si le nouveau-né pourrait être sensible à l'unité moraïque composée d'une **voyelle allongée**, Les résultats des études antérieures ont montré que leur perception du rythme de la parole serait basée plutôt sur la syllabe et plus probablement sur le noyau vocalique [Ber88]. Le nouveau-né pourrait donc être sensible à l'unité moraïque marquée par **la durée** d'une voyelle (/:/). Nous allons, dans cette expérience, observer si le nouveau-né est capable de discriminer cette unité parmi les mots bi-syllabiques, lorsque la différence porte seulement sur **la durée** de la première voyelle (l'allongement vocalique).

2.1. METHODE

Sujets : Tous les sujets sont des nouveau-nés français de 2-4 jours nés à la maternité Port Royal à Paris. 86 nouveau-nés ont participé à l'expérience. Parmi ces enfants, les données de 26 sujets ont été analysées. Ces derniers sont répartis en deux groupes : 12 pour le groupe contrôle et 14 pour le groupe expérimental (M. : 3370 g (poids), 68 heures (âge), 40 semaines (terme)).
Les enfants qui s'endormaient, pleuraient ou qui avaient des succions trop irrégulières ont été exclus de l'expérience.

Procédure : Nous utilisons la méthode basée sur la réponse de succion non-nutritive avec les phases de habituation et de déshabituation (la procédure SHA : Succions de Haute Amplitude) [Siq69; Eim71]. Nous observons, avec cette méthode, une différence significative de taux de succions de haute amplitude entre les enfants du groupe contrôle (avec les stimuli bi-moraïques) et du groupe expérimental, (avec les stimuli bi- et tri- moraïques). Si nous observons une augmentation des taux de succions significativement plus **élevée** chez les enfants du groupe expérimental que chez ceux du groupe contrôle, nous pouvons en déduire que les enfants du groupe expérimental discriminent les stimuli bi-moraïques de ceux tri-moraïques.

Stimuli : Les stimuli sont 5 mots japonais bi-syllabiques **bi**-moraïques et 5 mots japonais bi-syllabiques **tri**-moraïques :
Bi-moraïques : [rama], [sede], [itɨ], [koto], [musu]
vs
Tri-moraïques : [ra:ma], [se:de], [i:tɨ], [ko:to], [mu:su]
La variation des stimuli a été réduite en appliquant les règles suivantes :
1) Les stimuli ont tous un contour descendant :
Haut- Bas(-Bas) ; M. 254 Hz / 174 Hz.
2) A l'intérieur de chaque mot, la première et la deuxième voyelles sont identiques (ex. [koto]).
3) L'unité moraïque infra-syllabique est toujours insérée après la première voyelle de mots bi-syllabiques, bi-moraïques sans d'autres changements phonétiques.

2.2. RESULTATS

Une analyse de variance a été effectuée à partir des taux de succions sur les minutes de **Ligne de base** (aucun stimuli présentés), sur les **5 dernières minutes avant** le changement de phase (un seul type de stimuli) et les **4 minutes consécutives après** le changement de phase (mêmes stimuli pour le groupe contrôle, stimuli différents pour le groupe expérimental).
Nous avons ensuite observé le taux de succions sur **les 2 minutes avant et les 2 minutes après** le changement de phase.
Nous avons constaté une **augmentation significative de taux de succions** pour cette période pour le groupe contrôle : $F(1,10) = 8,41$, $p = ,0158$, + 5,8 HA (succions de Haute Amplitude), ainsi que pour le groupe expérimental : $F(1,12) = 8,64$, $p = ,0124$, +5,4 HA.

La différence d'augmentation de taux de succions pour ces deux groupes durant les deux minutes avant et les deux minutes après le changement de phase n'est donc pas **significative**.

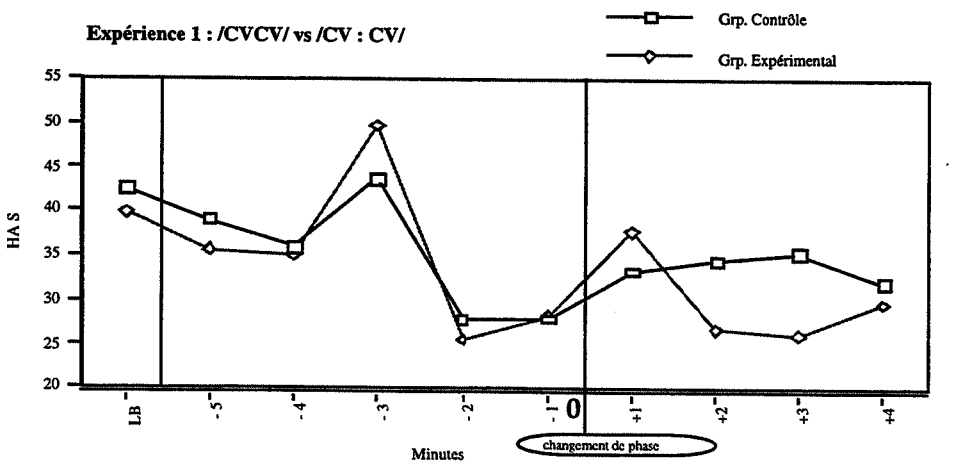


Figure 1: Moyenne du taux de succions de hautes amplitudes

3. EXPERIENCE 2

Cette deuxième expérience a pour but d'observer si le nouveau-né est capable de discriminer une autre sorte d'unité moraïque. Il s'agit d'une more constituée d'une consonne nasale /N/.
Contrairement à l'unité moraïque vocalique testée dans l'expérience 1, celle-ci est une unité **consonantique**. La présence de cette unité marque la différence entre une syllabe ouverte (CV) et une syllabe fermée (CVN).

3.1. METHODE

Sujets : Les critères de sélection de sujets sont identiques à l'expérience 1. Dans cette deuxième expérience, nous avons testé 80 nouveau-nés et analysé les données de 26 sujets. Ces derniers sont répartis en deux groupes : 12 pour le groupe contrôle et 14 pour le groupe expérimental (M. : 3348,50 g (poids), 74 heures (âge), 39 semaines (terme)).

procédure : Idem à l'expérience 1.

Stimuli : Les stimuli sont 5 mots japonais bi-syllabiques bi-moraïques et 5 mots japonais bi-syllabiques tri-moraïques qui ont été sélectionnés de la même façon que dans l'expérience 1 :

Bi-moraïques : [saka], [mede], [itɨ], [koto], [bugu]
vs
Tri-moraïques : [saNka], [meNde], [iNtɨ], [koNto], [buNgu]

3.2. RESULTATS

L'analyse de variance a été effectuée de la même manière que dans l'expérience 1.

Nous avons, de nouveau, observé une **augmentation significative de taux de succions** pour le groupe **contrôle** : $F(1,10) = 8,04$, $p = ,0177$, + 7,5 HA, ainsi que pour le groupe **expérimental** : $F(1,12) = 6,49$, $p = ,0256$, + 7,5 HA.

La différence d'augmentation de taux de succions pour ces deux groupes durant les deux minutes avant et les deux minutes après le changement de phase n'est donc pas **significative**.

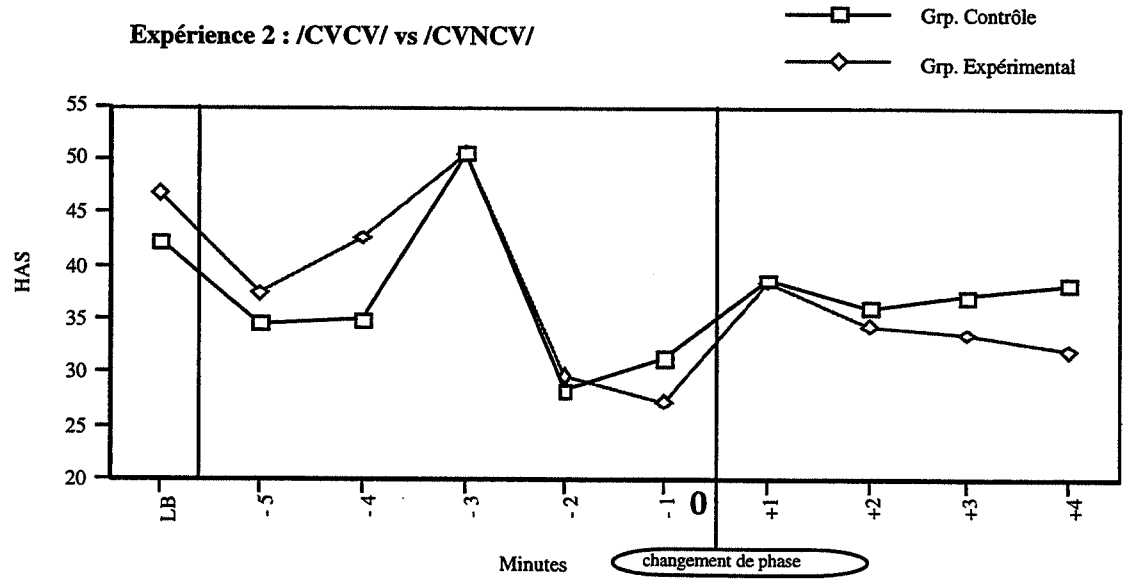


Figure 2 : Moyenne du taux de succions de hautes amplitudes

4. DISCUSSION GENERALE

Nous avons, dans les deux expériences, observé une augmentation significative de taux de succions non seulement chez les enfants du groupe expérimental à qui on présente les **deux différentes sortes de stimuli**, mais également chez les enfants du groupe contrôle à qui on présente **une seule sorte de stimuli**.

Les données obtenus ne montrent, en effet, aucune différence significative de taux de succions suivant les stimuli qui différencient le nombre de mores.

D'après les résultats obtenus dans nos deux expériences, nous pouvons conclure que **le nouveau-né français de 2-4 jours ne discrimine pas d'unités moraïques** insérées dans des mots japonais malgré le contexte extrêmement simplifié.

Ces résultats correspondent à ceux obtenus dans les expériences antérieures [Ber95]. Ils appuient, une nouvelle fois, l'hypothèse que l'unité de représentation à l'état initial serait plutôt basée sur l'unité syllabique et que l'enfant développerait ensuite une unité de représentation appropriée à sa langue maternelle.

5. BIBLIOGRAPHIE

[Ber88]
Bertoncini J., Bijeljac-Babic R., Jusczyk P., Kennedy L. et Mehler J. (1988), "An investigation of young infants' perceptual representations of speech sounds", Journal of Experimental Psychology General, vol. 117, pp. 21-33
[Ber95]
Bertoncini J., Floccia C., Nazzi T. et Mehler J. (1995), "Morae and syllables : rhythmical basis of speech representation in neonates", Langage and Speech, vol. 38 (4), pp. 311-329
[Bij93]
Bijeljac-Babic R., Bertoncini J. et Mehler J. (1993), "How do four-day old infants categorize multisyllabic utterances ?", Developmental Psychology, vol. 29 (4), pp. 711-721.
[Chr93]
Christophe A. (1993), "Rôle de la prosodie dans la segmentation en mots", Doctorat de l'Ecole des Hautes Etudes en Sciences Sociales.
[Chr94]
Christophe A., Dupoux E., Bertoncini J. et Mehler J. (1994), "Do infants perceive word boundaries ? An empirical study of the bootstrapping of lexical acquisition", J.Acoust.Soc.Am, vol. 95 (2), pp. 1-11.
[Cut86]
Cutler A., Mehler J., Norris J. et Segui J. (1986), "A langage specific comprehension strategy", Nature, vol. 304, pp. 159-160.
[Cut86]
Cutler A., Mehler J., Norris J. et Segui J. (1986), "The syllable's differing role in the segmentation of French and English", Journal of Memory and Language, vol. 25, pp. 385-400.
[Dec80]
De casper A. et Fifer W.P. (1980), "Of human bonding : Newborns prefer their mothers' voices", Science, vol. 208, pp. 1174-1176.
[Eim71]
Eimas P.D., Siqueland E.R., Jusczyk P. et Vigorito J. (1971), "Speech perception in infants", Science, vol. 171, pp. 303-306.
[Meh81]
Mehler J., Dommergues J., Frauenfelder U. et Segui J. (1981), "The syllable's role in speech segmentation", Journal of Verbal Learning and Verbal Behaviour, vol. 20, pp. 298-305.
[Meh86]
Mehler J., Lambertz G., Jusczyk P. et Amiel-Tison C. (1986), "Discrimination de la langue maternelle par le nouveau-né", Compte Rendu de l'Académie des Sciences de Paris, vol. 15, pp. 637- 640.
[Ota93]
Otake T., Hatano G., Cutler A. et Mehler J. (1993), "Mora or Syllable ? Speech segmentation in Japanese", Journal of Memory and Language, vol. 32, pp. 258- 278.
[Siq69]
Siqueland E.R. et De Lucia C.A. (1969), "Visual reinforcement of non-nutritive sucking in human infants", Science, vol. 165, pp. 1144-1146.

Production

Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM

P. Badin[1], L. Pouchoy[1], G. Bailly[1], M. Raybaudi[2], C. Segebarth[2], J.F. Lebas[3], M. Tiede[4], E. Vatikiotis-Bateson[4], Y. Tohkura[4]

[1] Institut de la Communication Parlée, UPRESA CNRS 5009, INPG - Université Stendhal
46 avenue Félix Viallet, 38031 Grenoble Cedex 01, France
Tél.: ++33 476 57 48 26 - Fax: ++33 476 57 47 10 – e-mail: (badin,bailly)@icp.inpg.fr
[2] Unité INSERM U438, Grenoble, France
[3] Service d'IRM, Centre Hospitalier Régional Universitaire, Grenoble, France
[4] Human Information Processing Res. Labs., ATR International, Kyoto, Japon

ABSTRACT

Based on a set of 3D vocal tract images obtained by MRI, a 3D statistical articulatory model has been built using guided principal component analysis. It constitutes an extension to the lateral dimension of the previously mid-sagittal model developped from a radiofilm taken on the same subject. The parameters of the 2D model have been found to be good predictors of the 3D shapes, for most configurations.

1. INTRODUCTION

Les modèles articulatoires constituent un moyen privilégié d'étudier les phénomènes de production de la parole, et en particulier leur contrôle. Cependant, les modèles traditionnels sont limités à la seule coupe médiosagittale du conduit vocal, ce qui conduit à un certain nombre de problèmes: (1) il est nécessaire d'inférer la fonction d'aire à partir de la coupe sagittale [Beau96], (2) les consonnes latérales qui présentent une occlusion dans la zone médiosagittale mais avec des canaux latéraux ouverts ne peuvent pas être traités par ces modèles, et (3) les modes acoustiques transverses qui se propagent à partir de 4-5 kHz ne peuvent pas être pris en compte. L'intérêt de disposer de modèles réellement tridimensionnels apparaît donc clairement. En outre, ces modèles ouvrent des possibilités d'accéder à la réalité virtuelle par l'élaboration de têtes parlantes audiovisuelles utiles pour la synthèse audiovisuelle, l'aide à l'apprentissage des langues, etc.

L'objectif de notre étude est donc le développement d'un modèle articulatoire tridimensionnel basé sur un corpus de données géométriques acquises par IRM pour un sujet. A notre connaissance, le seul modèle de ce type est celui développé par [Tie96]. Ce modèle était limité à des voyelles et n'était pas doté de paramètres de commande ayant une interprétation en termes articulatoires. La seule autre tentative de modélisation tridimensionnelle est celle de [Sto97], limitée à une section coronale dans la région palatale, sur des données obtenues par imagerie ultrasonique. Il existe par contre des modèles articulatoires tridimensionnels par éléments finis (cf. [Tri95]), qui sont très lourds à mettre en oeuvre, et pas nécessairement mieux maîtrisés du point de vue de leur contrôle.

Notre modèle constitue une extension à la troisième dimension d'un modèle articulatoire médiosagittal [Beau96] suivant un principe d'analyse en composantes principales guidée. Son originalité réside dans la compatibilité ascendante, c'est-à-dire que la réduction de ce nouveau modèle au plan médiosagittal est identique au modèle médiosagittal bidimensionnel de départ, avec en particulier les mêmes commandes articulatoires.

2. DONNÉES GÉOMÉTRIQUES 3D

2.1 Acquisition des données IRM

Pour chacune des configurations du corpus, 55 coupes perpendiculaires au plan sagittal ont été obtenues à l'aide du scanner IRM Philips GyroScan T10-NT de 1 Tesla disponible au CHU de Grenoble. Les coupes, d'épaisseur 3,6 mm, espacées de 4,0 mm, ont été réalisées en mode *Spin Echo*, et présentent une résolution finale de 1 mm/pixel. Elles sont groupées en trois piles de coupes parallèles, l'une coronale, la deuxième inclinée à 45° et la troisième axiale, ajustées de manière à couvrir l'ensemble du conduit vocal du sujet, tout en restant au maximum perpendiculaires à la ligne médiane du conduit (voir figure 1). L'ensemble des 55 coupes est acquis en 43 secondes, ce qui permet au sujet de soutenir artificiellement la configuration, soit en apnée complète, soit en expirant très faiblement en mode chuchoté. Notons que pendant le silence précédent / suivant la réalisation (très bruyante !) des coupes, le sujet devait prononcer la voyelle ou respectivement le début / la fin de la séquence VCV, de manière à fournir une référence du signal de parole correspondant.

2.2 Dépouillement des images

Le dépouillement vise à reconstituer les contours tridimensionnels du conduit vocal sous forme d'une série de contours planaires situés dans des plans perpendiculaires au plan médiosagittal et passant par les lignes de grille du modèle médiosagittal (cf. [Beau96]). Rappelons que ce système de grille comporte une partie centrale polaire fixe, et deux parties extrêmes linéaires, mais ajustées en longueur l'une entre la pointe de la langue et la limite postérieure du palais dur, et l'autre entre la racine de la

langue et les parties haute et basse du larynx. La première étape de dépouillement consiste à déterminer, pour chacune des coupes des trois piles, les contours du conduit vocal. Ceci est réalisé par simple seuillage, à l'aide du logiciel public de traitement d'image NIH-Image qui délivre le contour sous forme de points X/Y codant le contour par un code de Freeman.

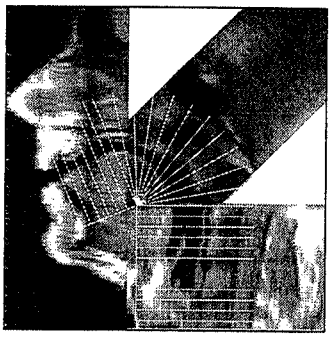


Figure1: Exemple de positionnement de la grille sur une coupe médiosagittale reconstituée à partir des trois piles de départ.

Une coupe médiosagittale est par ailleurs reconstituée à partir des trois piles de coupes à l'aide de NIH-Image pour chacune des articulations du corpus (voir figure 1). L'ensemble des contours déterminés dans l'étape précédente, et associés à la coupe médiosagittale reconstruite, sont calés, à l'aide d'une *roto-translation* (rotation suivie d'une translation), par rapport au repère fixe commun que constitue le tracé médiosagittal du palais dur, complété par le contour externe du conduit vocal, qui est relativement stable, obtnus par cinéradiographie. Les paramètres de la roto-translation nécessaire pour amener en superposition les deux contours médiosagittaux permettent par la suite de recalculer les coordonnées des contours dans le repère de référence du modèle sagittal. Enfin, sur la même coupe médiosagittale reconstituée, sont marqués les points qui correspondent à la pointe et à la racine de la langue, et aux parties haute et basse du larynx, permettant ainsi de préciser de manière complète l'ensemble de la grille.

L'étape suivante consiste à rééchantillonner les contours dans chacun des plans de coupe initiaux avec un nombre identique de points d'abcisses curvilignes réparties de manière équistante (51 semble un nombre approprié), et à les lisser à l'aide de filtres passe-bas de Butterworth appliqués à chacune des coordonnées. Les points de même indice de chaque contour sont ensuite associés en lignes tridimensionnelles, ou *fibres*, qui constituent un maillage du contour du conduit vocal. Finalement, les intersections de chacune de ces fibres avec les plans perpendiculaires au plan médiosagittal et passant par les lignes de grille du modèle médiosagittal sont calculées. Il en résulte des contours planaires en nombre égal au nombre de lignes de la grille. Une représentation de la surface correspondante est proposée à la figure 2.

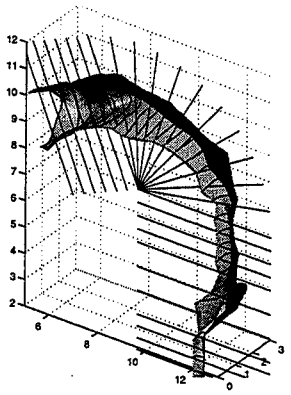


Figure 2: Exemple de représentation de la surface du conduit vocal pour la voyelle [a], avec superposition des contours planaires et de la grille de référence.

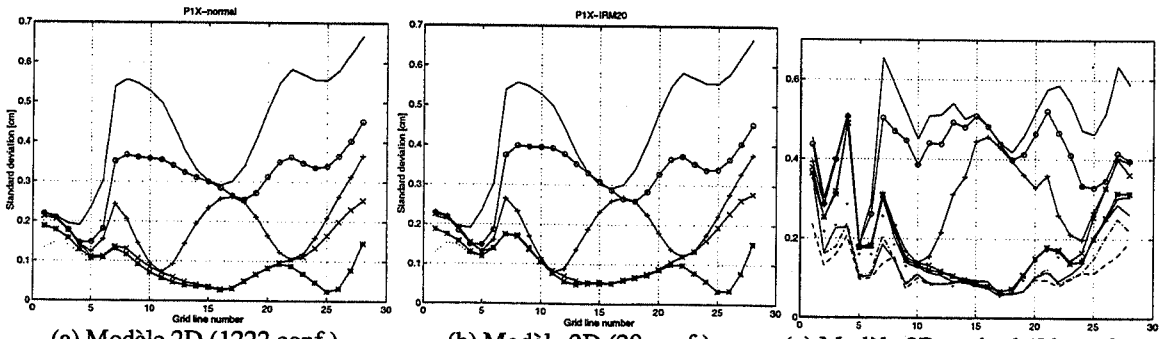
3. ACP GUIDÉE

3.1 Principe

Rappelons que l'ACP guidée (cf. [Beau96]) consiste à déterminer, de manière itérative, un ou plusieurs prédicteurs pour les résidus des variables auxquelles les les contributions des prédicteurs précédents ont déjà été soustraites. Cette technique, si elle ne conduit pas à une explication maximale de la variance par un nombre minimal de facteurs, permet par contre de choisir des facteurs interprétables de manière claire, ou mesurables, telle la hauteur de la mâchoire par exemple.

3.2 Effet de la taille du corpus

Rappelons par ailleurs que les modèles articulatoires statistiques dans le plan médiosagittal sont en général



(a) Modèle 2D (1222 conf.) (b) Modèle 2D (20 conf.) (c) Modèle 3D sagittal (20 conf.)

Figure 3: Ecarts-type (en cm) des abscisses sur la grille dans le plan médiosagittal (–) et de leurs résidus après soustraction des contributions de JH (O), TB (+), TD (×), TT (*), TA ()–, P1 (•), P2 (–), P3 (·-) et P4(--).

construits à partir d'au moins 500 à 1000 contours [Beau96]. Nous avons vérifié qu'en choisissant convenablement les échantillons, c'est-à-dire en sélectionnant seulement les cibles vocaliques et consonantiques du corpus de départ, on peut obtenir un modèle articulatoire qui approche les données de manière presque aussi satisfaisante que le modèle établi sur l'ensemble du corpus. Nous avons en particulier montré que les erreurs de reconstruction des données, calculées comme la racine carrée de la moyenne quadratique des erreurs d'abscisse sur l'ensemble des points de la grille et des 1222 images du corpus de contours médiosagittaux dont nous disposons [Beau96], étaient de 0.9 mm, 1.1 mm et 1.7 mm lorque le modèle était élaboré à partir respectivement de 1222, 20 et 8 configurations. Ceci est corroboré par la figure 3(a, b) qui montre les écarts-types des abscisses et de leurs résidus après soustraction des contributions des paramètres du modèle 2D: *hauteur de la mâchoire* JH, *corps* TB, *dos* TD, *pointe* TT et *avancée* TA de la langue [Beau96].

Le corpus d'images IRM que nous avons constitué pour le sujet PB comportait donc un nombre limité d'articulations: les 10 voyelles orales du français, et les consonnes [p t k f s ʃ ʀ l] soutenues, censées être produites dans les trois contextes symétriques [a i u], soit un total de 34 cibles. Dans l'état actuel, seules 20 configurations ont été dépouillées et utilisées.

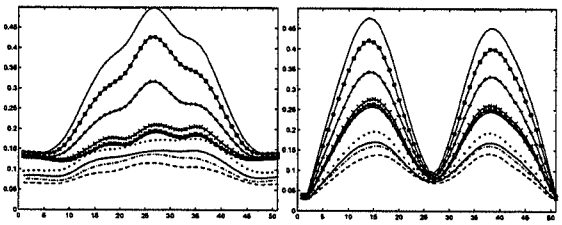
3.3 Analyse des données 3D

Les contours planaires obtenus à partir des images IRM sont ensuite traités par analyse en composantes principales guidée. Notons que, pour simplifier cette première approche d'une modélisation 3D du conduit vocal, nous limitons l'analyse à la partie du conduit vocal qui contient la langue, en excluant provisoirement les cavités en aval de l'apex. L'ensemble des variables à analyser est donc constitué des coordonnées X/Y de chacun des 51 points des 28 contours planaires correspondant à la grille, soit un total de 2856 variables. Dans chacun des plans de coupe, la coordonnée X correspond à la *dimension latérale* du contour, tandis que la coordonnée Y correspond à un déplacement dans la direction des droites de la grille de référence, que nous appellerons *dimension sagittale*.

Afin que le modèle 3D soit une extension du modèle 2D, nous imposons comme premiers prédicteurs de l'ensemble des 2856 variables les paramètres JH, TB, TD, TT et TA. Pour chaque articulation, ces paramètres sont obtenus par inversion à partir du contour intérieur 2D de la langue constitué par les points d'intersection des contours planaires avec le plan médiosagittal. Nous avons donc déterminé un par un, par régression linéaire, les coefficients de prédiction des coordonnées des contours liés à ces paramètres. Indiquons que le coefficient de qualité, qui exprime la variance des données expliquée par les prédicteurs JH, TB, TD, TT et TA est de 75 %. Les quatre premiers fateurs, P1, P2, P3 et P4, résultant de l'ACP classique que nous avons ensuite appliquée aux résidus de cette première analyse font monter l'explication de la variance à 94 % de la variance totale, mais aucun effet particulièrement interprétable n'émerge de manière forte de ces facteurs.

L'évolution des écarts-type résiduels moyennés sur l'ensemble des points d'une fibre des 1428 coordonnées *latérales* et *sagittales* en fonction des prédicteurs employés est illustrée à la figure 4 pour chacune des 51 fibres. On notera l'approximative symétrie en ce qui concerne la dimension sagittale ; les écarts-type pour les fibres proches du contour médiosagittal externe (proches des fibres 1 et 51) devraient être nuls : les valeurs non nulles constatées s'expliquent d'une part par des mouvements du velum présents même dans des articulations non nasales, et d'autre part par les bruits de mesure du palais dur et de la paroi arrière du pharynx. Notons d'autre part que la fibre 25 correspond au coutour médiosagittal interne, et donc au modèle 2D. La distribution des écarts-type des résidus en fonction de l'indice du contour pour cette fibre 25 est présentée à la figure 3c. On peut noter les très fortes similitudes avec les figures équivalentes du modèle 2D (figures 3a et 3b), ce qui confirme la cohérence entre les données 3D acquises par IRM et les données 2D acquises par cinéradiographie sur le même sujet. La distribution concernant la dimension latérale est aussi naturellement symétrique, et présente des zéros plus marqués au voisinage des fibres 1, 25 et 51 situées par construction dans le plan médiosagittal. Nous constatons que JH, TB, TD sont les principaux prédicteurs pour la dimension sagittale, tandis que JH, TB, TD et P1 sont les principaux prédicteurs pour la dimension latérale.

Les paramètres JH, TB, TD, TT et TA n'étant pas orthogonaux par construction, puisque issus de l'inversion du modèle 2D, nous avons cependant vérifié qu'ils ne sont que faiblement corrélés (voir table 1), à l'exception des paramètres de l'apex, ceci étant dû à l'abscence de contextes vocaliques [u] dans le corpus traité pour l'instant.



Dimension *sagittale* Dimension *latérale*

Figure 4: Ecarts-type résiduels moyennées sur l'ensemble des points d'une fibre en fonction du numéro de la fibre. Mêmes symboles qu'à la figure 3.

Table 1: Matrice de corrélation des paramètres mesurés

	JH	TB	TD	TT	TA
JH	1.000	0.126	0.087	0.052	-0.592
TB	0.126	1.000	0.017	0.364	0.289
TD	0.087	0.017	1.000	0.670	-0.310
TT	0.052	0.364	0.670	1.000	0.074
TA	-0.592	0.289	-0.310	0.074	1.000

4. Le modèle tridimensionnel

Nous avons finalement établi un modèle tridimensionnel du conduit vocal, qui prédit les coordonnées X/Y des contours planaires dans les différents plan de la grille comme des combinaisons linéaires des neufs paramètres de commande JH, TB, TD, TT, TA, P1, P2, P3 et P4. Nous avons ensuite analysé les effets des différentes commandes en mettant à zéro tous les paramètres sauf un, et en calcul la forme du conduit pour les valeurs –3 et +3 du paramètre analysé. La figure 5 illustre les résultats pour JH, TB, TD et TT. Nous avons par ailleurs confirmé que les paramètres P*i* n'ont pas d'influence majeure sur la forme du conduit.

5. Discussion

Cette étude préliminaire montre que les paramètres de commande du modèle 2D médiosagittal suffisent à commander le modèle 3D de manière très satisfaisante. Il faut cependant modérer cette conclusion par le fait que notre analyse n'incluait pas les cavités en aval de la pointe de la langue, en particulier la cavité sublinguale et les lèvres, et que le corpus ne comprenait pas les consonnes latérales qui nécessitent à l'évidence des commandes supplémentaires pour gérer les canaux latéraux en présence d'une occlusion dans le plan médiosagittal.

Cette étude a donc montré la faisabilité d'un modèle statistique tridimensionnel du conduit vocal basé sur les principes fructueux appliqués aux modèles médiosagittaux. Il est évident que de nombreux travaux seront nécessaires avant d'aboutir à un modèle réellement exploitable, à la fois au niveau du traitement des images pour réduire les bruit de mesure (liés en particulier au difficile positionnement de la grille en l'absence de repères osseux fiables) et au niveau de la méthode de représentation des contours concernant la pointe de la langue, la luette et l'épiglotte.

Remerciements

Ce travail a été mené dans le cadre du Programme de Recherche "Une Tête Parlante Virtuelle: Données et modèles en production de parole" financé par l'Agence Rhône-Alpes pour les Sciences Sociales et Humaines. Nous sommes reconnaissants pour leur aide à nos collègues d'ATR, du CHRUG (en particulier C. Rubin), et de l'ICP (en particulier L. Apostol, A. Arnal, C. Lecourtier et L. Revéret).

Bibliographie

[Beau96] Beautemps D., Badin P., Bailly G., Galván A., & Laboissière R. (1996), "Evaluation of an articulatory-acoustic model based on a reference subject", 4th Speech Production Seminar / ETRW, pp. 45-48.

[ElM98] El Masri S., Pelorson X., Saguet P., & Badin P. (en révision), "Development of the Transmission Line Matrix method in acoustics. Applications to higher modes in the vocal tract and other complex ducts", International Journal of Numerical Modelling.

[Sto97] Stone M., Goldstein M.H., & Zhang Y. (1997), "Principal component analysis of cross sections of tongue shapes in vowel production", Speech Communication, Vol. 22, pp. 173-184.

[Tie96] Tiede M., Yehia H., & Vatikiotis-Bateson E. (1996), "A shape-based approach to vocal tract area function estimation", 4th Speech Production Seminar / ETRW, pp. 41-44.

[Tri95] Wilhelms-Tricarico R. (1995), "Physiological modeling of speech production: Methods for modeling soft-tissues articulator", J. Acoust. Soc. Am., Vol. 97, pp. 3085-3098.

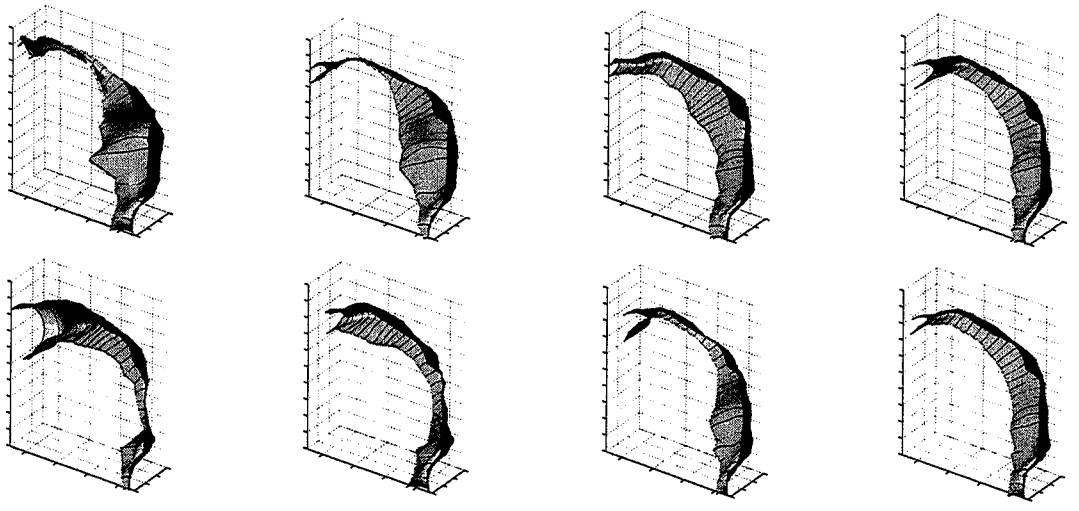


Figure 5: Nomogrammes du modèle 3D : les 9 facteurs sont nuls, sauf le facteur mentionné qui prend la valeur –3 (haut) et +3 (bas).

Contribution de la mâchoire à la géométrie de la langue dans les modèles articulatoires statistiques

Gérard Bailly, Pierre Badin & Anne Vilain

Institut de la Communication Parlée — INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél. : ++33 04 76 57 47 11 - Fax : ++33 04 76 57 47 10
e-mail : {bailly,badin,vilain}@icp.inpg.fr - http ://www.icp.inpg.fr/

RÉSUMÉ

Linear component articulatory models [Mae79, Mae88, Gab94] are built using an iterative substraction of linear predictors of the vocal tract geometry. In this paper we consider the contribution of jaw displacement to tongue and lips movements using sets of cineradiographic data from three different speakers. We show that linear prediction overestimates this contribution by capturing not only the intrinsic mechanical jaw-tongue coupling but also the synergetic control observed in the corpus. We then propose a substraction of the jaw contribution which do not affect the performance of the model in terms of data prediction.

1. INTRODUCTION

Les modèles articulatoires permettent de contrôler la forme du conduit vocal à l'aide d'un nombre minimal de paramètres [BMP94]. Dans la plupart des modèles, la forme contrôlée directement est la fonction sagittale. On peut distinguer trois types de modèle : (a) les modèles géométriques [Mer73] considèrent la forme du conduit vocal comme un ensemble de figures géométriques simples (droites, arcs). Certains points ou angles caractéristiques sont utilisés comme paramètres de contrôle ; (b) les modèles biomécaniques [Per69, WT95, SLO97] considèrent la structure musculo-skelettale. Les paramètres de contrôle sont alors les niveaux d'activation musculaire ; (c) les modèles statistiques [HLG77, Mae79, Gab94] cherchent à prédire les formes du conduit vocal d'un locuteur prononçant un corpus limité. Les formes sont obtenues par imagerie médicale dynamique (cinéradiographies, ultrasons ...) ou statique (images IRM ...). Les paramètres de contrôle sont ici les prédicteurs (en général linéaires) de ces formes caractérisées par un nombre constant de points. Ceci est rendu possible par l'usage d'une grille de lecture [Mae88] ou l'échantillonage d'une abscisse curviligne [KH96].

Faire parler ces modèles suppose de pouvoir accéder à des paramètres de contrôle les plus indépendants possibles et auxquels une interprétation neuroanatomique puisse être donnée. L'un des articulateurs dont les degrés de liberté sont les mieux connus, les plus facile à mesurer et qui sont à la base du développement du langage, est certainement la mâchoire. L'oscillation mandibulaire permet d'expliquer les associations préférentielles de l'ontogénèse : l'hypothèse "pure frame" [DM95] suppose l'émergence des premières consonnes par une élévation de la mâchoire avec une langue inerte correspondant à diverses configurations vocaliques (voir simulations dans [BBVB95]).

Il est donc crucial de disposer d'un couplage réaliste entre la mâchoire et les organes qu'elle porte. Si ce couplage est réalisé classiquement dans les modèles géométriques par l'hypothèse "mécaniste" (voir par exemple le modèle de Lindblom et Sundberg [LS71]) décrite à la section 3, les modèles statistiques laissent généralement ce soin à une prédiction linéaire. Nous allons soumettre ces modèles à des données cinéradiographiques de trois sujets enregistrées à l'hopital Schiltigheim de Strasbourg : B1X [Mae88], P1X [BGB+95] et J1X [Vil97].

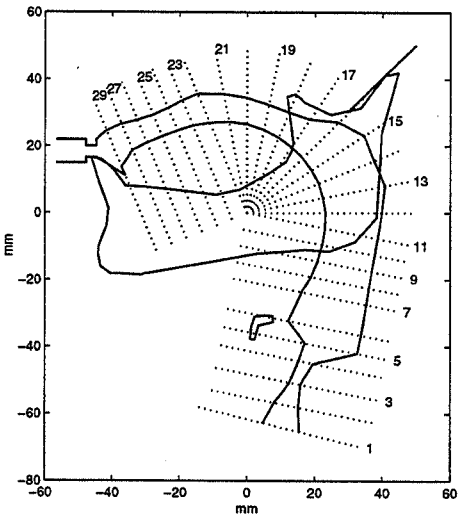
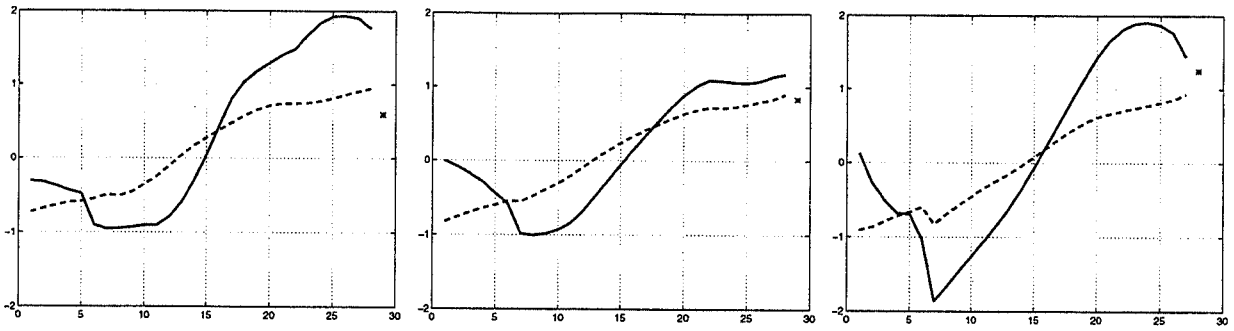


FIG. 1: Grille utilisée pour l'analyse/synthèse de P1X superposée au contour synthétique "neutre". Les droites de 7 à 22 de la grille sont fixes pour chaque locuteur. Les lignes 1..6 sont positionnées par rapport à la hauteur du larynx et les lignes 23..29 par rapport à la pointe de la langue.

2. PRÉDICTION LINÉAIRE DES FORMES DE LANGUES

La grille est définie dans [Gab94]. Cette grille possède des parties fixes, liées aux dents et à la région vélaire, et des parties mobiles liées au larynx, au bout de la langue et aux



(a) B1X (b) P1X (c) J1X

FIG. 2: Pente de la régression linéaire reliant le déplacement de la mâchoire aux points d'intersection de deux langues avec la grille pour les trois corpus. En traits pleins, la langue réelle, en traits pointillés, la langue moyenne supposée rigide. L'étoile figure le coefficient de régression pour l'écartement labial.

lèvres. Un exemple de grille est donné figure 1. Le point 15 de la grille correspond grossièrement à l'extrémité du vélum.

La figure 2 montre, pour les trois sujets, les relations existant entre le déplacement de l'incisive inférieure et les déplacements des points d'intersection du contour sagittal de la langue avec la grille. On constate tout d'abord que ces "fonctions de couplage" ont une grande cohérence : quand la mâchoire se ferme, (a) la partie avant de la langue (points de grille supérieurs à 18) monte, (b) la partie arrière de la langue (points de grille entre 7 et 15) s'avance et (c) le bas pharynx (points de grille entre 1 et 6) s'élargit. Les amplitudes des déplacements de la partie avant et arrière de la langue croissent de manière symétrique par rapport au point 15 de la grille.

On remarque cependant que cette régression linéaire opérée sur les données brutes met en évidence des déplacements dans la partie avant de la langue plus importants que les déplacements de la mâchoire elle-même alors que la grille est bien positionnée suivant l'axe des mouvements et que les points de la langue sont bien circonscrits à l'intérieur du lieu décrit par une rotation/translation simple de la mâchoire. Le rapport entre déplacements atteint 1,95 pour B1X et 1,9 pour J1X. Pour la partie arrière de la langue de J1X, la prédiction linéaire prédit de même des rapports atteignant 1,85. Seul, P1X montre une courbe qui évolue pratiquement entre les bornes physiquement réaliste [-1,+1]. On observe en outre que la pente de régression du déplacement des lèvres est voisine de l'unité. C'est d'ailleurs cette valeur qui est utilisée dans les modèles initiaux issus de ces données et cités en introduction.

3. HYPOTHÈSE "MÉCANISTE"

Ces rapports moyens ne peuvent s'expliquer physiquement qu'en faisant l'hypothèse "irréaliste" que les déplacements de la mâchoire et ceux de la langue sont couplés par un système dynamique largement sur-amorti. Dans l'interprétation quasi-statique des mouvements faite dans les modèles statistiques, ils donnent une grande importance aux mouvements de la mâchoire dans l'exécution des mouvements avant-arrière de la langue (voir figure 3(a)) et, de manière symétrique, un faible découplage langue-mâchoire dans les procédures d'inversion acousti-articulatoire. Nous avons ci-après caractérisé les rapports de déplacements induits par un couplage mécanique de la rotation/translation mandibulaire avec une langue moyenne supposée rigide.

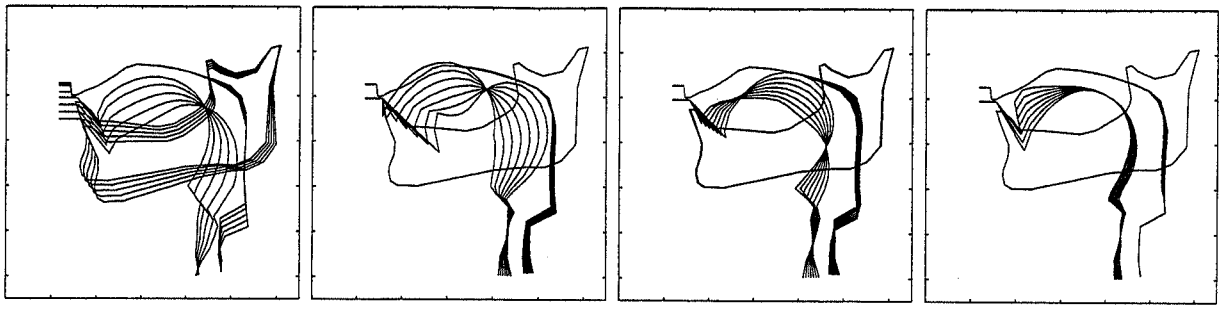
3.1. *Langue rigide attachée à la mâchoire*

La figure 2 montre également pour les trois sujets, les relations existant entre le déplacement de l'incisive inférieure et les déplacements de la langue moyenne. Pour ceci, on considère un ensemble mâchoire/langue moyenne rigide. Le mouvement de deux points sur la mâchoire, extraits des tracés cinéradiographiques, permet de calculer pour chaque contour la rotation/translation subie par cet ensemble. L'intersection avec la grille de la langue moyenne ainsi déplacée est calculée et le déplacement induit sur la langue moyenne ainsi déterminé. La figure 2 montre (courbes en pointillés) que, pour la partie linguale, les rapports entre le déplacement de l'incisive inférieure et les points de la langue moyenne présentent une allure similaire à celle des pentes de régression commentées à la section 2. Comme attendu, les deux courbes sont dans un rapport de 2 :1 environ et les valeurs maximales absolues dans notre cas ne dépassent pas 1,0. Pour la partie du bas-pharynx cependant, le modèle de couplage mécanique prévoit une influence grandissante de l'articulation mandibulaire de l'épiglotte au larynx alors que la régression laisse cette zone de plus en plus insensible aux mouvements de mâchoire.

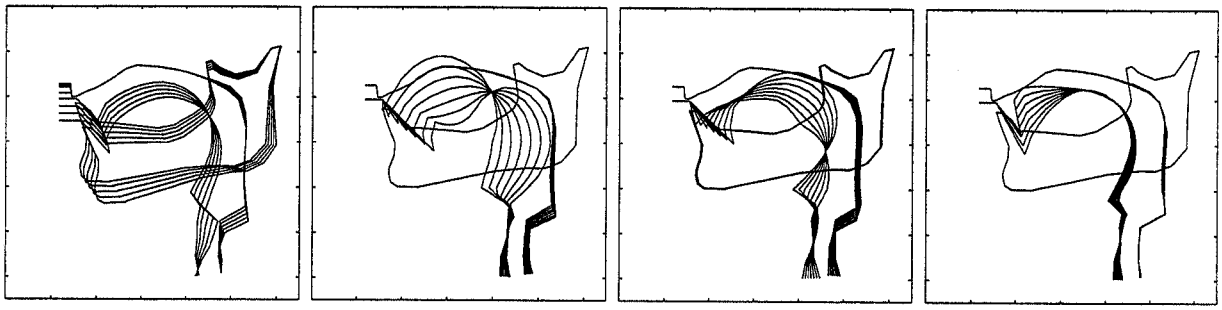
3.2. *Modèle statistique composite*

Ces données nous ont suggéré une composition des deux modèles visant à minimiser l'influence de la mâchoire sur les déplacements des organes portés : le modèle composite que nous proposons choisit pour chaque point du contour labial le coefficient d'amplitude minimale en valeur absolue. Cette fusion de modèles revient à conserver le modèle initial pour la partie bas-pharynx et à opter pour le modèle "mécanique" pour la partie linguale. Ces modèles sont remarquablement continus dans la zone de transition entre bas-pharynx et langue.

Nous avons vérifié que la poursuite de la procédure de



(a) régression linéaire



(b) modèle composite

FIG. 3: Nomogrammes articulatoires pour J1X obtenus par deux méthodes de soustraction de la composante mâchoire. Les mouvements à partir de la configuration moyenne sont induits par, de gauche à droite : la mâchoire, le corps, le dos et l'élévation de la pointe de la langue.

soustraction de composantes linéaires n'entraînait pas de pertes de prédiction des contours sagittaux. Rappelons que les composantes linéaires extraites sont dans l'ordre après la mâchoire (MA) : le dos (TD), le corps (TB), l'élévation (TT) et l'avancée (TA) de la pointe de la langue. Le paramètre TA permet de contrôler finement la partie non expliquée par les paramètres MA, TD et TB de la distance entre les droites 23 et 29 de la figure 1. La figure 4 compare, pour chaque locuteur, l'évolution de la variance expliquée en chaque point de la grille et en fonction du nombre de composantes linéaires considérées, ceci pour deux modèles : (a) le modèle classique de régression linéaire (RL) et (b) le modèle composite (C). Notons tout d'abord que B1X a une très faible dynamique mandibulaire : l'incidence du prédicteur de couplage langue/mâchoire est donc réduite. Sur les deux autres locuteurs, l'évolution comparée de la variance expliquée suggère quelques commentaires : (a) comme attendu, la composante mâchoire de C n'explique qu'environ la moitié de la variance expliquée par celle de RL, (b) cet écart de variance est rattrapé par les deux paramètres suivants : corps et dos de langue, (c) TB rattrape pratiquement toute la variance sauf une petite portion de la partie centrale, domaine d'action principal de TD.

4. CONCLUSIONS

En prenant appui sur les bases de données cinéradiographiques de trois locuteurs (deux hommes : P1X et J1X et une femme : B1X), nous avons proposé une correction de la première étape de soustraction de composantes linéaires dans les modèles articulatoires statistiques. Le modèle composite proposé ne produit pas une prédiction sous-optimale des contours sagittaux des corpus originaux. Cette correction devrait permettre de mieux caractériser la contribution de la mâchoire aux mouvements des effecteurs finaux de l'articulation et ainsi mieux comprendre et modéliser la négociation entre canaux vocaliques et consonantiques du langage articulé.

5. REMERCIEMENTS

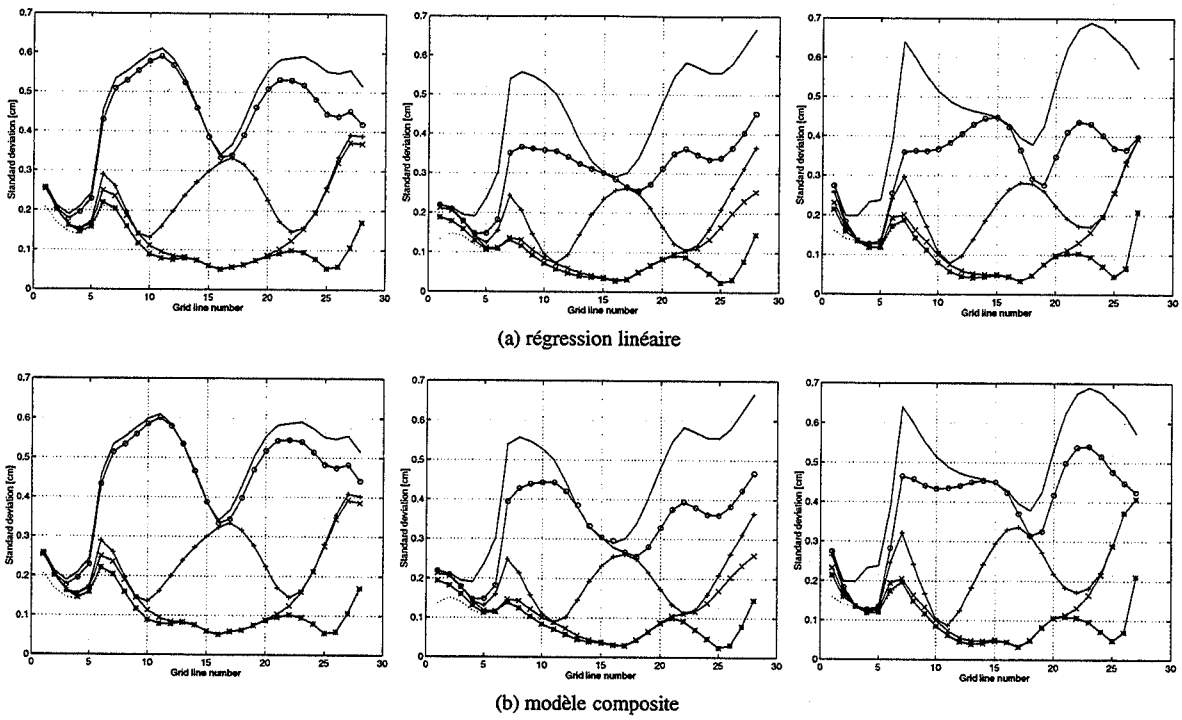
Nous tenons à remercier Shinji Maeda et nos collègues de l'Institut de Phonétique de Strasbourg, plus particulièrement Gilbert Brock pour leur collaboration dans l'obtention des données.

RÉFÉRENCES

[BBVB95] Gérard Bailly, Louis-Jean Boë, Nathalie Vallée, and Pierre Badin. Articulatori-acoustic prototypes for speech production. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 1913–1916, Madrid - Spain, 1995.

[BGB+95] Pierre Badin, Bernard Gabioud, Denis Beautemps, Tahar Lallouache, Gérard Bailly, Shinji Maeda, Jean-Pierre Zerling, and Gilbert Brock. Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model. In *International Congress on Acoustics*, pages 349–352, Trondheim - Norway, 1995.

[BMP94] Louis-Jean Boë, Shinji Maeda, and Pas-



(a) régression linéaire

(b) modèle composite

FIG. 4: Evolution comparée de la variance expliquée en chaque point de la grille et en fonction du nombre de composantes linéaires considérées pour le contour sagittal de la langue pour les deux méthodes d'estimation de la composante mâchoire. De haut en bas dans chaque dessin : la mâchoire (o), TB (+), TD (x) et TT ($*$). De gauche à droite : B1X, P1X et J1X.

cal Perrier. La modélisation articulatoire :un demi-siècle d'évolution entre fonctionnel, physique et biomécanique. In *Journées d'Etudes sur la Parole*, pages 41–54, Trégastel–France, 1994.

[DM95] Barbara L. Davis and Peter F. MacNeilage. The articulatory basis of babbling. *Journal of Speech and Hearing Research*, 38 :1199–1211, 1995.

[Gab94] Bernard Gabioud. Articulatory models in speech synthesis. In Eric Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 215–230. John Wiley and Sons, Chichester, 1994.

[HLG77] Richard Harshman, Peter Ladefoged, and Louis Goldstein. Factor analysis of tongue shapes. *Journal of the Acoustical Society of America*, 62 :693–707, 1977.

[KH96] Tokihiko Kaburagi and Masaaki Honda. A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes. *Journal of the Acoustical Society of America*, 99(5) :3154–3170, 1996.

[LS71] Bjorn Lindblom and J. Sundberg. Acoustic consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50 :1166–1179, 1971.

[Mae79] Shinji Maeda. An articulatory model of the tongue based on a statistical analysis. *Journal of the Acoustical Society of America*, 65 :S22, 1979.

[Mae88] Shinji Maeda. Improved articulatory model. *Journal of the Acoustical Society of America*, 81(S1) :S146, 1988.

[Mer73] Paul Mermelstein. An articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53 :1070–1082, 1973.

[Per69] Joseph Perkell. *Physiology of Speech Production*. MIT Press, Cambridge, MA, 1969.

[SLO97] Vittorio Sanguineti, Rafaël Laboissière, and David Ostry. An integrated model of the biomechanics and neural control of the tongue, jaw, hyoid and larynx system. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4, pages 2023–2026, Rhodes - Greece, 1997.

[Vil97] Anne Vilain. Un nouveau modèle articulatoire pour la synthèse et le contrôle robotique de la parole : Gentiane. Rapport de DEA Sciences du Langage, Université Stendhal – Grenoble, 1997. sous la direction de Christian Abry et Pierre Badin.

[WT95] Reiner Wilhelms-Tricarico. Physiological modeling of speech production : Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, 5 :3085–3098, 1995.

Articulation consonantique en début de constituants prosodiques de différents niveaux.

Cécile FOUGERON

ILPGA-Paris III, URA1027

19 rue des Bernardins, 75005 Paris

e-mail : fougeron@humnet.ucla.edu

ABSTRACT

In this paper, I show that the articulation of a consonant is affected by its position in a prosodic constituant in French. Linguopalatal contact is compared for various consonants placed in four different prosodic positions: Intonational Phrase initial, Accentual Phrase initial, Word initial and Syllable initial. For the two speakers studied, consonants initial in higher prosodic domains have more linguopalatal contact than consonants initial in lower domains. This effect appears clearly in the articulation of the stops /n, t, k/ and the lateral /l/, but less clearly for the fricative /s/. These results suggest that some articulatory properties (here lingual constriction) can reflect, at the production level, the prosodic structure of an utterance, by marking the initial boundary of constituants at different prosodic levels.

1. Introduction

L'objectif de cette étude est de voir si l'organisation prosodique d'un énoncé est marquée, au niveau de la production, dans l'articulation des segments.

La majeure partie des études sur le sujet porte sur les variations dans la production des segments en fonction de leur position dans des constituants prosodiques inférieurs, comme la Syllabe ou le Mot. Ces études s'inscrivent dans un cadre de recherche sur le marquage des frontières lexicales dans le signal de parole et sur l'identification d'unités phonétiques correspondantes au Mot. Par exemple, il a été observé que le degré de constriction de la langue sur le palais [ex. Kea95], l'élévation du vélum [ex. Ben77], ou l'ouverture de la glotte [Coo91] sont plus importants pour les segments placés en début vs. en milieu ou fin de Mot. A ces variations spatiales de l'articulation s'ajoutent des variations temporelles qui consistent en un allongement des segments initiaux dans le Mot [Oll73].

Les recherches sur le marquage des frontières entre constituants supérieurs au Mot se sont concentrées sur les caractéristiques des segments en position finale. Les variations les plus connues sont celles associées aux variations temporelles des syllabes en fin de constituants. Ces syllabes sont allongées par rapport aux syllabes non-finales et cet allongement est proportionnel au degré de la frontière prosodique suivante (i.e. à la hauteur prosodique du constituant) [ex. Wig91]. Ainsi, l'allongement final (associé aux indices intonatifs) permet de délimiter des constituants prosodiques supérieurs au Mot. De plus, il reflète la hiérarchie des constituants en distinguant plusieurs niveaux de frontières.

Plus récemment, il a été montré que l'articulation des segments initiaux varie aussi en fonction du degré de la frontière prosodique qui les précède (et donc de la hauteur du constituant qu'ils débutent). Par exemple, l'aspiration des consonnes aspirées en anglais [Pie92] et en coréen [jun93], ou la constriction linguopalatale de /n/ en anglais [fou97] sont plus importantes pour les consonnes initiales dans des constituants de niveaux supérieurs par rapport à des consonnes initiales dans des constituants inférieurs. Ces variations articulatoires et acoustiques permettent de distinguer les frontières initiales de constituant à plusieurs niveaux prosodiques. Elles permettent aussi de distinguer dans un constituant les consonnes initiales des consonnes médiales (celles-ci étant les consonnes initiales dans un constituant inférieur).

L'étude que je présente ici s'inscrit dans ce cadre de recherche. Je cherche à savoir (1) si la segmentation prosodique d'un énoncé influence l'articulation des segments placés en début de constituant (à droite de la frontière) et (2) si cet effet reflète la hiérarchie des constituants prosodiques en marquant le début de constituants à différents niveaux prosodiques en français.

2. Méthode

Cinq consonnes, variant selon leurs lieux et modes d'articulation, sont étudiées : /t, n, k, l, s/. Elles sont encadrées par les voyelles ouvertes /a_a/ et /õ_õ/ pour /t/. Les consonnes sont placées à quatre positions prosodiques différentes : en début de Groupe Intonatif (GIi), de Groupe Accentuel (GAi), de Mot lexical (Mi), de Syllabe (Si) [fou98]. Pour la consonne /k/, la position Si n'est pas examinée. Les phrases cadres utilisées sont conçues pour obtenir un phrasé particulier dans lequel le segment test sera placé à la position prosodique souhaitée. Cette position est vérifiée à posteriori selon les critères prosodiques suivants : présence/absence de pause précédente (+, Ø), allongement de la voyelle précédente (++, +, 0), fonction démarcative du ton précédent (frontière majeure (%%), mineure (%), non démarcatif (Ø)). Tous les segments tests portent un ton bas. Le

tableau 1 illustre ce codage par position et présente les stimuli formés pour la consonne /n/.

Tableau 1	pause	allongement V1	ton frontière
1- GIi :	+	++	%%
La pauvre Tata. Nadia et Paul n'arriveront que demain.			
2- GAi :	Ø	+	%
Tonton, Tata, Nadia et Paul arriveront demain ...			
3- Mi :	Ø	0	Ø
Tonton-Paul et Tata-Nadia arriveront demain par le train.			
4- Si :	Ø	0	Ø
Tonton-Paul et Tata Annabelle arriveront demain ...			

Le degré de contact linguopalatal est mesuré à l'aide du Kay Palatometer avec une fréquence d'échantillonnage de 100 Hz et avec un palais artificiel couvert de 96 électrodes. Le signal acoustique correspondant est enregistré simultanément à une fréquence de 12800 Hz. Quinze à vingt répétitions de chaque stimuli ont été enregistrées par deux locuteurs parisiens (A et l'auteur B). Le degré (%) et la distribution du contact linguopalatal sont examinés au point où le contact est maximal pendant la constriction consonantique. L'articulation des consonnes en fonction de leur position prosodique est comparée à l'aide d'un test ANOVA (1 facteur "position") et des tests post-hoc Ficher PLSD (à 95%). Une position est considérée comme statistiquement différente des autres positions si elle se différencie de toutes les autres positions examinées.

3. Résultats

3.1 Consonnes occlusives /n, t, k/ :

Pour les deux locuteurs, le degré de contact linguopalatal des trois occlusives varie en fonction de la position de la consonne. La figure 1 illustre cette variation pour le locuteur A. Le degré de contact linguopalatal des consonnes tend à augmenter progressivement de la position prosodique la plus basse (Si) à la position la plus élevée (GIi). Pour la consonne /n/, le degré de contact linguopalatal permet de différencier de manière significative les quatre positions prosodiques étudiées. Pour les deux locuteurs, les /n/ en début de S ont moins de contacts que les /n/ en début de M, ceux-ci ont moins de contacts que les /n/ en début de GA, et ces derniers ont moins de contacts que les /n/ en début de GI.

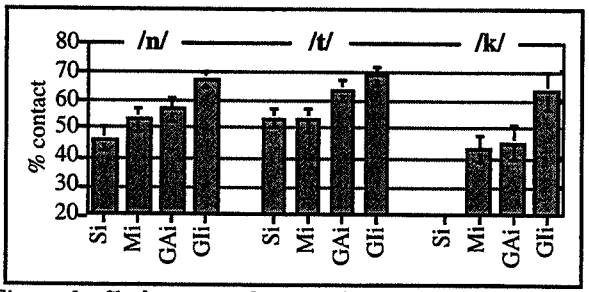


Figure 1 : % de contact linguopalatal pour /n/, /t/, et /k/ aux positions prosodiques examinées, locuteur A.

Pour la consonne /t/, seules 3 positions se distinguent. Pour le locuteur A (fig.1), le nombre de contact augmente des positions Si ou Mi à la position GAi, puis à la position GIi. Pour le locuteur B, le nombre de contact augmente de la position Si à GAi à GIi. La position Mi ne se distingue pas des positions Si ou GAi.

Pour la consonne /k/ seules les positions Mi, GAi et GIi sont comparées. Pour la consonne vélaire, le contact linguopalatal observé est limité à la surface du palais artificiel, i.e. au palais dur. Pour le locuteur A (fig. 1), les /k/ en début de M et de GA ont moins de contacts que les /k/ en début de GI. La différence entre les débuts de M et de GA (p=0,05) est à la limite de la significativité. Pour le locuteur B, les trois positions se distinguent par une augmentation progressive de contact.

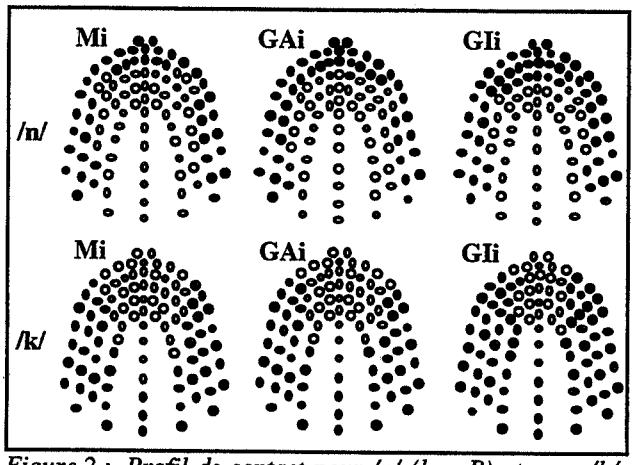


Figure 2 : Profil de contact pour /n/ (loc. B) et pour /k/ (loc A). Électrodes contactées représentées en noir.

Pour les trois occlusives, l'augmentation de contact se traduit par un élargissement progressif de la surface de l'occlusion, vers l'arrière du palais pour /t/ et /n/ et vers l'avant du palais pour /k/. Ceci est illustré sur la figure 2 qui présente un exemple de profil de contact aux positions Mi, GAi et GIi pour /n/ et /k/.

3.2 Consonne latérale /l/ :

Trois régions sont définies d'après les profils observés pour caractériser l'articulation de /l/ (fig.3). La région centrale (grisée) s'étend jusqu'à la zone pré-palatale et sur les cotés après les prémolaires. Les régions latérales (noire et blanche) couvrent les trois rangées latérales du palais au niveau des molaires.

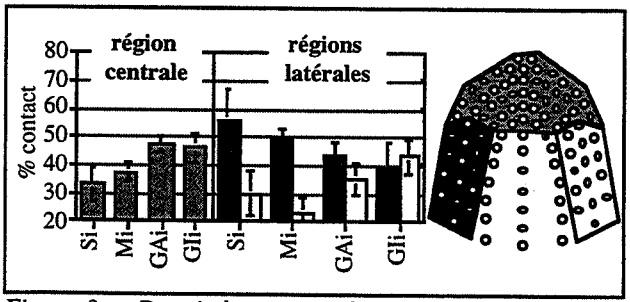


Figure 3 : Degré de contact dans les régions centrale (gris), gauche (noir) et droite (blanc) pour /l/. Loc. A.

Pour les deux locuteurs, le degré de contact au niveau de l'occlusion centrale distingue un niveau sous-lexical (Si),

un niveau lexical (Mi) et un niveau post-lexical (GAi, et GIi) avec une augmentation progressive du contact antérieur (fig. 3, loc.A). Sur les cotés, on observe pour les deux locuteurs, une augmentation du contact sur le coté de l'ouverture latérale (droit pour le loc. A) et une diminution du contact du coté de l'encrage de la langue (gauche pour le loc. A). Cette variation dans l'équilibre des contacts droits et gauches résulte en une diminution de l'asymétrie des contacts latéraux dans les positions prosodiques supérieures. Ceci apparaît si l'on compare la différence entre les barres noires et blanches sur la figure 3. La comparaison des positions prosodiques à l'aide d'un index d'asymétrie [(gauche - droite) / total] permet de distinguer les positions Si-Mi vs. GAi vs. GIi pour le locuteur A et Si-Mi vs. GAi-GIi pour le locuteur B.

3.3 Consonne fricative /s/ :

L'articulation de la fricative /s/ est observée sur la région antérieure du palais où se situe le chenal fricatif. Cette région (fig. 4) s'étend du milieu de la face interne des incisives à l'angle convexe des alvéoles et, sur les cotés, jusqu'aux prémolaires.

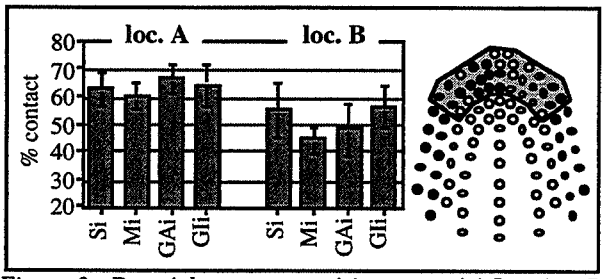


Figure 3 : Degré de contact antérieur pour /s/. Loc A et B.

Pour les deux locuteurs, le degré de contact linguopalatal dans cette région antérieure varie en fonction de la position prosodique mais les distinctions entre positions sont moins claires que celles observées pour les occlusives. Si on exclu la position Si, on a pour le locuteur A une augmentation du contact de la position Mi aux positions GAi-GIi. Pour le locuteur B, le contact augmente des positions Mi-GAi à la position GIi. En début de Syllabe, le degré de contact linguopalatal varie en fonction du locuteur : il est similaire aux positions GIi pour le locuteur B, et aux positions Mi-GIi pour le locuteur A. Les deux locuteurs n'ont donc en commun que l'augmentation de contact du début de Mot au début de Groupe Intonatif.

3.4 Variations acoustiques :

Dans cette section, il s'agit de voir si l'influence de la position prosodique observée sur le plan articulatoire, peut être observée sur le plan acoustique. Les variations articulatoires étudiées sont des variations de constriction linguale. Des variations dans la quantité de débit d'air nasal ont aussi été observées pour la nasale /n/ [voir fou98]. Il est fort probable que la position prosodique affecte plusieurs articulations, voir tout le système de production. Il n'est donc pas possible de prédire quelles seront les conséquences acoustiques précises des variations articulatoires en fonction de la position prosodique. Je n'ai examiné que quelques unes des caractéristiques acoustiques majeures des consonnes. Je présenterai ici les variations, pour les sonnantes /n/ et /l/, de la durée acoustique de la tenue consonantique et de l'énergie acoustique relative des consonnes à celle de la voyelle suivante (V2). Les résultats sont présentés dans le tableau suivant. La position prosodique influence les deux indices acoustiques analysés. Pourtant, aucun de ces indices n'apparaît comme "le" corrélât acoustique de la variation articulatoire observée.

	loc. A	loc. B
durée /n/	GAi>*Mi>*Si>*GIi	GAi >* Si, Mi, GIi
/l/	GAi>* Si,Mi,GIi; Mi>* GIi	GAi >* Si, Mi, GIi
énergie V2-/n/	Mi, GAi <* Si, GIi	Si, Mi <* GAi <* GIi
V2-/l/	Mi >* Si, GAi, GIi	Mi, GAi <* Si <* GIi

La durée des consonnes initiales ne suit que partiellement la hiérarchie des positions prosodiques. Seule la position GAi se distingue nettement des autres positions par un allongement de la consonne. Les consonnes en position GIi (précédées d'une pause) ont un degré de contact linguopalatal plus important que dans les positions inférieures, mais cette variation articulatoire ne s'accompagne pas d'un allongement de la durée acoustique.

Les variations dans l'énergie des consonnes sonnantes ne reflètent que très peu les variations articulatoires observées. La distinction la plus fréquente est une augmentation de la différence en terme d'énergie entre la consonne et la voyelle suivante en position GIi par rapport à Mi. Dans ces cas, la consonne devient plus saillante en terme d'énergie dans la syllabe CV et son caractère consonantique est mis en relief. Pourtant cette tendance est inversée dans le cas de /l/ pour le locuteur A.

4. Discussion et Conclusion

Comme il a été observé au niveau du Mot ou de la Syllabe, les consonnes en position initiale dans un constituant prosodique de niveau supérieur sont marquées sur le plan articulatoire. Le degré et la répartition du contact linguopalatal des consonnes permet de distinguer plusieurs niveaux prosodiques. Le nombre et la nature de ces niveaux dépendent des locuteurs et des consonnes étudiées. Pourtant, la tendance reste constante : plus la position prosodique s'élève dans la hiérarchie prosodique, plus le contact linguopalatal de la consonne est important. Le tableau suivant résume les distinctions observées :

	loc. A	loc. B
/n/	Si<*Mi<*GAi<*GIi	Si<*Mi<*GAi<*GIi
/t/	Si-Mi<*GAi<*GIi	Si<*GAi<*GIi
/k/	Mi-GAi<*GIi	Mi<*GAi<*GIi
/l/	Si<*Mi<*GAi-GIi	Si<*Mi<*GAi-GIi
asymétrie	Si-Mi>*GAi>*GIi	Si-Mi>*GAi-GIi
/s/	Mi<*GAi-GIi	Mi-GAi<*GIi-Si

En général, au moins 3 positions prosodiques sur 4 (2/3 pour /k/) se distinguent par une variation de contact. La distinction la plus consistante se fait entre les positions

extrêmes, Si-Mi et GIi. Les distinctions fréquentes entre les positions Mi et GAi indiquent que les variations observées ne reposent pas uniquement sur une distinction entre les consonnes précédées d'une pause (GIi) et celles non-précédée d'une pause (Si, Mi, GAi).

L'effet de la position prosodique apparaît nettement dans la production des occlusives et de la latérale observées. Par contre, la consonne fricative semble moins affectée par les variations de positions prosodiques. Ce résultat n'est pas surprenant et confirme la moindre variabilité des fricatives pour qui une trop forte modification du degré de constriction aurait une conséquence articulatoire et acoustique dramatique.

Les variations observées pour les occlusives et la latérale consistent en un élargissement de la constriction au niveau de l'occlusion linguale (antérieure pour /t, n, l/, postérieure pour /k/) et en un rétrécissement du chenal latéral pour /l/. Ces modifications suggèrent un raffermissement de la constriction de la langue contre le palais aux positions prosodiques supérieures. Ces variations sont similaires à celles décrites communément sous les termes de "renforcement articulatoire", "d'augmentation de l'effort articulatoire", de "fortition". Ces termes ont été utilisés pour caractériser, par exemple, les différences entre consonnes initiales et finales [Mar55], entre consonnes fortes et faibles [Mal70] ou entre parole renforcée et normale [Str63]. Bien que ce terme reste à définir physiologiquement, on pourrait parler "renforcement articulatoire" au début des constituants prosodiques de niveau supérieur [voir fou98 pour discussion].

Le marquage des frontières prosodiques est indiscutablement multiparamétrique. Les indices portés par les segments finaux (allongement, f0) sont nombreux et sont importants pour la segmentation prosodique. Cette étude montre que la structuration prosodique d'un énoncé est aussi marquée par l'articulation des consonnes en début de constituant. Les informations portées par les segments initiaux peuvent donc également contribuer à la segmentation prosodique. L'examen des variations acoustiques n'a pas mis à jour "un" corrélât acoustique qui, seul, permettrait de distinguer autant de positions prosodiques que ne le permettent les variations articulatoires observées. Pourtant, même en l'absence de corrélât acoustique précis, un renforcement articulatoire en début de constituant peut avoir une fonction linguistique. Il est connu qu'au niveau du Mot, les consonnes initiales présentent, tant du point de vue diachronique que synchronique, une résistance particulière aux phénomènes de réduction phonétique et phonologique. Par exemple, les consonnes initiales de Mot se sont maintenues du latin au français alors que les consonnes non-initiales ont été réduites jusqu'à parfois disparaître. D'autre part, les processus d'assimilation, de lénition ou de coarticulation affectent moins fréquemment les segments en début de Mot. On peut alors faire l'hypothèse qu'un renforcement articulatoire progressif des consonnes initiales à différents niveaux prosodiques, pourrait aussi avoir une fonction "préservatrice". Les consonnes "renforcées" aux frontières clefs dans l'organisation prosodique de l'énoncé seraient alors plus résistantes aux influences réductrices de la parole.

Références

[Ben77] Benguerel A-P. (1977) "Velar coarticulation in French: a fiberscopic study" *J. of Phonetics* 5, 149-58.

[Coo91] Cooper A. (1991) "Laryngeal and oral gestures in English /p, t, k/," *XIIth ICPhS*, 2, 50-53

[fou97] Fougeron C. & Keating P. (1997) "Articulatory strengthening at edges of prosodic domain", *JASA* 106, 3728-40

[fou98] Fougeron C. (en préparation) "Variations rticulatoires en fonction de la position dans un constituant prosodique en français" Thèse. Paris III.

[jun93] Jun S-A. (1993). *The phonetics and phonology of Korean prosody*. Ph.D diss, Ohio State University.

[Kea95] Keating P. (1995) "Effect of prosodic position on /t, d/ tongue/palate contact", *XIIIth ICPhS*.

[Mal70] Malécot, A. (1970) "The lenis-fortis opposition: its physiological parameters", *JASA* 47, 1588-1592

[Mar55] Martinet A. (1955) Economie des changements phonétiques. Francke, Berne.

[Oll73] Oller D. (1975) "The effect of position in utterance on speech segment duration in English", *JASA* 54, 1235-1247.

[Pie92] Pierrehumbert J. & Talkin D. (1992) "Lenition of /h/ and glottal stop". In *Papers in Lab. Phonology II: Gesture, Segment, Prosody,* Docherty & Ladd (eds). Cambridge U. Press, Cambridge UK, Chap. 4, 90-117.

[Str63] Straka G. (1963) "La division des sons du langage en voyelles et consonnes: peut-elle être justifiée?," *Travaux de Linguistique et de Littérature*, Université de. Strasbourg, I, 17-99.

[Wig91] Wightman C. et al. (1991) "Segmental durations in the vicinity of prosodic phrase boundaries", *JASA* 92, 1707-1717

Inversion acoustique articulatoire par une méthode variationnelle

Yves Laprie et Bruno Mathieu
LORIA
BP 239
54506 Vandœuvre-lès-Nancy, FRANCE
laprie,mathieu@loria.fr

RÉSUMÉ

This paper presents a novel approach to recovering articulatory trajectories from the speech signal using a variational calculus method and Maeda's articulatory model. The acoustic-to-articulatory mapping is generally assessed by a double criterion: the acoustic proximity of results to acoustic data and the smoothness of articulatory trajectories.

Most of the existing methods are unable to exploit the two criteria simultaneously or at least at the same level. On the other hand, our variational calculus approach combines the two criteria simultaneously and ensures the global acoustic and articulatory consistency without further optimization. This method gives rise to an iterative process which optimizes a startup solution given by an improved lookup algorithm. Codebooks generated with an articulatory model show nonuniform sampling of the acoustic space due to nonlinearities of the acoustic-to-articulatory mapping. We therefore designed an improved lookup algorithm building realistic articulatory trajectories which are not necessarily defined throughout the speech signal.

1. INTRODUCTION

L'estimation de la forme du conduit vocal à partir du signal de parole suscite de nombreux efforts à cause des applications potentielles tant en synthèse et codage de la parole qu'en reconnaissance automatique. La plupart des travaux sur l'inversion acoustique-articulatoire repose sur l'utilisation de modèles articulatoires qui assurent une approximation réaliste du conduit vocal avec un nombre limité de paramètres.

Il n'existe pas de relation bi-univoque entre les domaines acoustique et articulatoire et l'estimation de paramètres articulatoires à partir du signal nécessite donc des méthodes de tabulation ou de neuromimétisme. Mais quelle que soit la méthode, elle doit conduire à des paramètres articulatoires variant lentement dans le temps et qui produisent des spectres aussi proches que possible de ceux de la parole originale. Cette exigence correspond à la satisfaction de deux critères : la proximité avec les données acoustiques (par exemple le spectrogramme) et la régularité des trajectoires articulatoires. Ces deux critères sont nécessaires pour mesurer la qualité du passage de l'acoustique à l'articulatoire.

Généralement les méthodes existantes ne prennent pas en compte ces critères au même niveau, ou au moins favorise l'un des critères au détriment de l'autre. Dans le cas des méthodes par tabulation, par exemple, la distance acoustique est utilisée pour récupérer les formes du conduit vocal. Ensuite, des contraintes portant sur l'évolution des paramètres articulatoires assurent la régularité des trajectoires articulatoires. Souvent d'ailleurs, une optimisation finale améliore à chaque instant la proximité acoustique avec le signal d'origine.

Ces méthodes d'inversion sont souvent développées dans le but de réaliser le codage articulatoire sans que le réalisme des trajectoires articulatoires récupérées soit la préoccupation principale. En revanche, cette préoccupation devient primordiale lorsqu'il s'agit de trouver les commandes des articulateurs ou d'évaluer la position et le degré de la constriction principale du conduit vocal. Dans ce cas, les méthodes développées (par exemple [4]) reposent sur une approximation du modèle articulatoire (modèle « forward » et sur un contrôleur articulatoire (utilisant le calcul d'un pseudo-inverse du jacobien) capable de déformer le conduit pour atteindre une cible acoustique donnée.

Dans cet article nous proposons une nouvelle méthode pour combiner les deux critères que nous venons d'évoquer. Cette méthode repose sur le calcul variationnel qui donne lieu à un processus itératif. Ce processus commence avec une solution initiale grossière et construit une suite de trajectoires articulatoires qui optimise une fonction de coût combinant la distance acoustique et l'évolution des paramètres articulatoires. Cette approche présente trois avantages vis-à-vis de la plupart des autres méthodes :

- la nature continue des trajectoires formantiques et la cohérence acoustique et articulatoire sont directement prises en compte par la méthode elle même, et ne nécessitent donc pas d'optimisation supplémentaire.
- le comportement acoustique du modèle articulatoire est directement intégré à la méthode et ne nécessite donc pas d'apprentissage préalable de fonctions de sensibilité des formants par rapport aux paramètres articulatoires. Ces fonctions sont évaluées directement pendant l'inversion et correspondent exactement à la configuration articulatoire courante.
- il est possible d'examiner la transformation des solutions initiales lors de l'optimisation. Cela est particulièrement intéressant pour étudier les effets de com-

pensation en considérant des solutions initiales qui correspondent à plusieurs positions relatives des articulateurs.

2. PRÉSENTATION DE LA MÉTHODE

Nous avons retenu le modèle articulatoire de Maeda [7] qui décrit la forme du conduit vocal grâce à sept paramètres. Ces paramètres sont des fonctions du temps $\alpha(t) = (\alpha_1(t) \dots \alpha_i(t) \dots \alpha_7(t)), t \in [t_i, t_f]$. Les trajectoires formantiques extraites de la parole $f_j(t)$, $1 \leq j \leq 3$ sont les données du problème. Celles générées par la simulation acoustique sont notées $F_j(\alpha(t))$ $(1 \leq j \leq 3)$. La fonction de coût pour évaluer l'inversion comporte deux termes :

- $\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)))^2$ qui exprime la proximité entre les trajectoires formantiques observées $f_j(t)$ et $F_j(\alpha(t))$ celles générées par le modèle
- $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$ qui exprime la vitesse d'évolution des paramètres articulatoires. Pour pénaliser les efforts articulatoires trop intenses et éviter au conduit vocal d'atteindre des positions trop éloignées de la position de repos nous avons ajouté le terme $\sum_{i=1}^{7} k_i \alpha_i^2(t)$ qui représente l'énergie potentielle.

La fonction de coût à minimiser est donc de la forme suivante:

$$I = \int_{t_i}^{t_f} \sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)))^2 dt + \lambda \int_{t_i}^{t_f} \sum_{i=1}^{7} m_i \alpha_i'^2(t) dt + \beta \int_{t_i}^{t_f} \sum_{i=1}^{7} k_i \alpha_i^2(t) dt \quad (1)$$

où t_i and t_f définissent l'intervalle de temps sur lequel l'inversion a lieu, λ et β expriment le compromis entre la vitesse d'évolution des paramètres articulatoires, leur distance à la position de repos et la distance acoustique. m_i est la pseudo-masse du ième articulateur, et k_i sa constante de raideur. L'équation (1) peut s'écrire

$$I = \int_{t_i}^{t_f} \Phi(\alpha(t), \alpha'(t), t) dt$$

Le calcul variationnel [3] peut être utilisé pour minimiser I et les équations d'Euler-Lagrange expriment l'annulation de la dérivée de I par rapport à chacun des α_i. Ces équations, qui sont une condition nécessaire pour assurer qu'un minimum de I a été atteint, s'écrivent

$$\begin{cases} \frac{\partial \Phi}{\partial \alpha_1} - \frac{d}{dt}\frac{\partial \Phi}{\partial \alpha_1'} = 0 \\ \dots \\ \frac{\partial \Phi}{\partial \alpha_7} - \frac{d}{dt}\frac{\partial \Phi}{\partial \alpha_7'} = 0 \end{cases} \quad (2)$$

En remplaçant Φ par sa définition dans l'équation (1), chacune des équations d'Euler-Lagrange devient :

$$\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t)) \frac{\partial F_j}{\partial \alpha_i} + \beta k_i \alpha_i(t) - \lambda m_i \alpha_i''(t) = 0 \quad i = 1 \dots 7 \quad (3)$$

où $\alpha_i''(t)$ est la dérivée seconde par rapport au temps de $\alpha_i(t)$. À partir de maintenant nous ne considérerons qu'une seule des équations du système (2) pour ne pas alourdir les notations. Nous supposons que nous disposons d'une estimation grossière des trajectoires articulatoires α_i (cf. § 3). Nous pouvons alors définir un processus itératif $\alpha_i^\tau(t)$ tel que

$$\lim_{\tau \to \infty} \alpha_i^\tau(t) = \alpha_i(t)$$

(où $\alpha_i^{\tau=0}(t)$ est la solution initiale) en utilisant l'équation d'évolution associée

$$\gamma \frac{\partial \alpha_i^\tau}{\partial \tau} + \beta k_i \alpha_i^\tau - \lambda m_i \alpha_i^{\tau\prime\prime} = -\sum_{j=1}^{3}(f_j(t) - F_j(\alpha^\tau(t))) \frac{\partial F_j}{\partial \alpha_i^\tau} \quad (4)$$

$\frac{\partial \alpha_i^\tau}{\partial \tau}$ représente l'évolution du paramètre α_i au cours du processus itératif et γ un paramètre pour contrôler la vitesse d'évolution. Une solution à l'équation statique (3) est atteinte quand le terme $\gamma \frac{\partial \alpha_i^\tau}{\partial \tau}$ s'annule.

Par souci de simplicité nous avons fixé m et k à 1 mais nous reviendrons sur ce choix à la fin de cet article.

Notons $\alpha^\tau = (\alpha_{i,0}^\tau, \dots \alpha_{i,k}^\tau, \dots \alpha_{i,N}^\tau)$ la représentation discrète de $\alpha_i(t)$, $\alpha_{i,k}^\tau$ représente la valeur de α_i^τ au temps discret $t = t_i + k\frac{t_f - t_i}{N}$ lors de l'itération τ. Puisque la résolution de l'équation (4) pour α_i est indépendante des autres trajectoires articulatoires, $\alpha_{i,k}^\tau$ est noté α_k^τ pour simplifier les notations.

Notons $(f_0, \dots f_k, \dots f_N)$ la trajectoire formantique observée et $(F_0, \dots F_k, \dots F_N)$ la trajectoire formantique générée par la simulation acoustique. L'approximation aux différences finies de la dérivée $\alpha''(t)$ conduit à l'équation suivante

$$\gamma(\alpha_k^\tau - \alpha_k^{\tau-1}) + \beta\alpha_k^\tau - \lambda(\alpha_{k+1}^\tau - 2\alpha_k^\tau + \alpha_{k-1}^\tau) = -\sum_{j=1}^{3}(f_{j,k} - F_{j,k}) \left.\frac{\partial F_j}{\partial \alpha}\right|_{\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau} \quad (5)$$

où τ représente l'itération en cours de traitement et k le temps discret. Le terme de dérivation $\left.\frac{\partial F_j}{\partial \alpha}\right|_{\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau}$ est calculé pour le paramètre α_i au point $(\alpha_{1,k}^\tau \dots \alpha_{7,k}^\tau)$ et intègre le comportement acoustique du modèle par rapport à l'évolution des paramètres articulatoires.

Il faut ajouter des conditions aux limites pour assurer que l'équation (5) ait une solution unique. Puisque nous n'imposons aucune contrainte sur la position des extrémités de $\alpha(t)$

$$\alpha'(0) = \alpha'(N) = 0$$

sont donc les conditions aux limites naturelles. Soit B la matrice $(N+1) \times (N+1)$

$$B = \begin{bmatrix} \gamma+\beta+\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & \gamma+\beta+2\lambda & -\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\lambda & \gamma+\beta+2\lambda & -\lambda \\ 0 & \cdots & 0 & -\lambda & \gamma+\beta+\lambda \end{bmatrix}$$

$$\alpha^\tau = (\alpha_0^\tau, \dots \alpha_k^\tau, \dots \alpha_N^\tau)^T$$

$$c^\tau = \begin{bmatrix} \gamma\alpha_0^{\tau-1} - \sum_{j=1}^{3}(f_{j,0} - F_{j,0}) \frac{\partial F_j}{\partial \alpha} \\ \gamma\alpha_1^{\tau-1} - \sum_{j=1}^{3}(f_{j,1} - F_{j,1}) \frac{\partial F_j}{\partial \alpha} \\ \dots \\ \gamma\alpha_N^{\tau-1} - \sum_{j=1}^{3}(f_{j,N} - F_{j,N}) \frac{\partial F_j}{\partial \alpha} \end{bmatrix}$$

L'équation Eq. (5) peut s'écrire sous la forme matricielle suivante $B\alpha^\tau = c^\tau$

La résolution de l'équation (2) conduit à un processus itératif. α^τ est calculé à chaque itération pour chacun des sept paramètres articulatoires α_i. Pour garantir qu'il est possible d'atteindre un minimum de l'équation (1) if faut disposer d'une bonne solution initiale. Pour cela nous avons utilisé une méthode issue de l'approche par tabulation. La solution initiale est ensuite transformée de telle manière que l'équation (1) soit minimisée.

3. Choix d'une solution initiale

La fonction de coût définie par l'équation (1) n'est pas convexe et il faut donc une solution initiale située à proximité d'un minimum de I. Lors d'expériences préliminaires nous avons constaté qu'il est possible d'approcher très précisément les trajectoires formantiques dès que le dictionnaire de formes utilisé par la tabulation est très vaste. Cela n'assure cependant pas que les trajectoires articulatoires soient lisses et pertinentes, au moins lorsque les contraintes de la programmation dynamique ne sont pas trop fortes. Plusieurs raisons expliquent pourquoi les formes articulatoires à t et $t-1$ peuvent être très éloignées l'une de l'autre. La première est que l'échantillonnage du dictionnaire n'est pas suffisamment fin. Il n'y a donc pas de point d'entrée proche de la forme trouvée à $t-1$ qui correspond aux formants observés. Cela peut se produire à proximité d'une non-linéarité de la relation acoustique articulatoire (c'est-à-dire lorsqu'une petite variation des paramètres articulatoires donne une grande variation acoustique). La seconde raison est que le modèle articulatoire n'est pas capable de produire une forme de conduit identique à celle réalisée par le locuteur. Les formants sont donc atteints au moyen d'un effet compensatoire injustifié ce qui donne lieu à un point éloigné du reste de la trajectoire articulatoire. La seconde raison est d'autant plus importante qu'une adaptation profonde du modèle articulatoire n'est pas possible.

Nous préférons donc une solution partielle (qui n'est pas nécessairement définie sur tout l'intervalle de l'inversion) plutôt qu'une solution définie partout mais irréaliste et difficile à optimiser. L'approche variationnelle présente l'avantage qu'une solution partielle complétée par interpolation linéaire pour les points manquants est transformée conformément au comportement acoustique du modèle articulatoire. Cela garantit donc sa pertinence.

Nous nous sommes inspirés d'un algorithme de lissage non-linéaire proposé par Ney [9] pour concevoir un algorithme d'inversion par tabulation. Le but de l'algorithme de Ney est de lisser une courbe donnée sous la forme d'un ensemble de points donc certains très éloignés de la courbe attendue. Ney propose d'utiliser la programmation dynamique pour trouver les points qu'il faut conserver. Dans notre cas cela conduit au choix d'une séquence de formes, cette séquence n'étant donc pas forcément définie à chaque instant.

4. Expérimentations

Nous avons commencé par adapter le modèle à composantes linéaires de Maeda à notre locuteur en réalisant une série d'images IRM [8]. L'adaptation a porté sur les facteurs de taille de la bouche et du pharynx ainsi que sur la partie fixe du conduit vocal. Les composantes linéaires elles-mêmes n'ont pas été modifiées car elles résultent de l'analyse statistique d'un grand nombre d'images aux rayons X. Malgré l'adaptation il n'est pas possible d'assurer que le modèle soit capable d'approcher parfaitement les fonctions d'aire que notre sujet produit.

Les dictionnaires de formes utilisés par une méthode d'inversion par tabulation peuvent être construits soit par un échantillonnage aléatoire de l'espace des paramètres articulatoires, soit à partir de formes de base [6]. Les formes de base correspondent aux configurations articulatoires de voyelles prononcées isolément et les formes du dictionnaires sont obtenues en échantillonnant toutes les trajectoires reliant deux formes de base. La dispersion des appariements entre les domaines acoustique et articulatoire est nettement plus faible avec le dictionnaire construit à partir des formes de base (voir [8]). Nous avons pourtant retenu la méthode d'échantillonnage aléatoire parce que certaines des voyelles prononcées par notre locuteur avaient été nettement centralisées (sans doute à cause du bruit produit par la machine d'imagerie IRM) et ne pouvaient donc pas être utilisées comme formes de base. Nous avons construit un dictionnaire contenant 300000 formes articulatoires ; les formes ne pouvant pas correspondre à une fonction d'aire vocalique réaliste ont été éliminées grâce aux contraintes proposées par Boë et al. [2].

Les figures Fig. 1 et Fig. 2 montrent les trajectoires formantiques générées à partir des paramètres articulatoires récupérés à partir des formants qui ont été extraits de la parole, et trois des paramètres articulatoires (position de la mâchoire, position de la langue et ouverture des lèvres pour la figure Fig. 1, position de la mâchoire, position et forme de la langue pour la figure Fig. 2). Les formants ont été extraits avec notre algorithme de suivi [5].

Dans certains cas, en particulier quand la voyelle /a/ suit une voyelle fermée (comme par exemple dans la figure Fig. 2), nous avons observé un effet compensatoire exagéré entre la position de la mâchoire et l'ouverture des lèvres. Cela est dû à la minimisation du terme $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$ qui favorise la minimisation du paramètres des lèvres plutôt que celui de la mâchoire. Cela vient du fait que les pseudo-masses et les constantes de raideur ont été fixées à 1 ce qui n'est pas réaliste puisque l'inertie de la mâchoire est beaucoup plus grande que celle des lèvres. Les pseudo-masses et les constantes de raideur devraient être fixées sur la base de mesures physiologiques pour éviter ces effets compensatoires injustifiés. L'autre solution adoptée dans [1] consiste à apprendre les commandes motrices du conduit vocal qui contrôlent les variables articulatoires (notamment la position et le degré de la constriction principale). Cette solution évite l'utilisation d'un dictionnaire de formes mais réduit l'espace des solutions possibles.

En dépit d'améliorations nécessaires concernant les para-

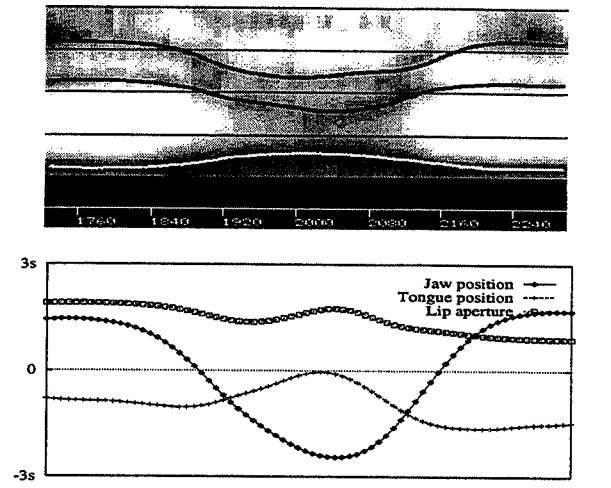


FIG. 1 – *Trajectoires formantiques générées à partir des trajectoires articulatoires extraites du signal /iai/ et trois des paramètres articulatoires représentant les écarts par rapport à la position neutre.*

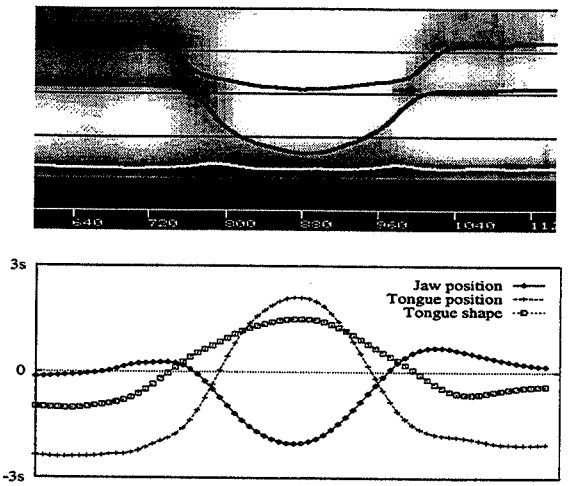


FIG. 2 – *Trajectoires formantiques générées à partir des trajectoires articulatoires extraites du signal /iui/ et trois des paramètres articulatoires représentant les écarts par rapport à la position neutre.*

mètres m_i et k_i notre méthode montre des qualité de flexibilité qui en font un bon outil pour étudier en particulier les effets de compensation. Notre algorithme de détermination des trajectoires initiales se révèle particulièrement efficace puisqu'il élimine les points erronés qui conduisent souvent à des trajectoires articulatoires très chaotiques.

5. CONCLUSION

Notre méthode scinde complètement le choix de la solution initiale et sa déformation pilotée par le modèle articulatoire. Il est donc facile d'étudier plusieurs solutions initiales et de comparer les solutions auxquelles elles conduisent. Cela présente un double intérêt : **(i)** celui de pouvoir comparer les valeurs du critère que définit l'équation Eq. 1 et d'éliminer certains minima locaux, **(ii)** et celui de pouvoir évaluer les phénomènes de compensation en choisissant les solutions initiales en fonction de l'effet à étudier.

Contrairement à d'autres techniques d'inversion acoustique-articulatoire notre méthode possède intrinsèquement un comportement dynamique car la solution initiale se déforme selon les fonctions de sensibilité des formants par rapport aux paramètres articulatoires. Cette approche a l'avantage de ne pas imposer de contraintes trop fortes sur les solutions initiales.

Le système actuel intègre le comportement acoustique du modèle articulatoire de Maeda. Mais le comportement dynamique du conduit vocal est en partie contrôlé par les pseudo-masses et les constantes de raideur que nous avons fixées arbitrairement à 1. Nous sommes en train d'étudier des méthodes permettant d'optimiser ces paramètres, par exemple des méthodes de validation croisée [10].

Remerciements : Nous remercions Shinji Maeda pour avoir mis à notre disposition son modèle articulatoire et Marie-Odile Berger pour de fructueuses discussions.

RÉFÉRENCES

[1] Bailly (G.), Abry (C.), Laboissièrre (R.), Perrier (P.) et Schwartz (J.-L.). – Inversion and speech recognition. *In : Signal processing VI: Theories and Applications*, éd. par Vandewalle (J.), Boite (R.), Mooner (M.) et Osterlinck (A.). pp. 159–164. – Brussels, Belgium, August 1992.

[2] Boë (L.-J.), Perrier (P.) et Bailly (G.). – The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics*, vol. 20, 1992, pp. 27–38.

[3] Bonvalet (M.). – *Les principes variationnels*. – Masson, Paris, 1993.

[4] Laboissière (R.) et Galvàn (A.). – Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences. *In : Proceedings ICPhS*, pp. 358–361. – Stockholm, August 1995.

[5] Laprie (Y.) et Berger (M.-O.). – Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, vol. 19, n4, October 1996, pp. 255–270.

[6] Larar (J. N.), Schroeter (J.) et Sondhi (M. M.). – Vector quantization of the articulatory space. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, n12, December 1988, pp. 1812–1818.

[7] Maeda (S.). – Un modèle articulatoire de la langue avec des composantes linéaires. *In : Actes 10èmes Journées d'Etude sur la Parole*, pp. 152–162. – Grenoble, Mai 1979.

[8] Mathieu (B.) et Laprie (Y.). – Adaptation of Maeda's model for acoustic to articulatory inversion. *In : Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 2015–2018. – Rhodes, Greece, September, 1997.

[9] Ney (H.). – A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, vol. 5, n2, March 1983, pp. 163–173.

[10] Richards (H. B.), Bridle (J. S.), Hunt (M. J.) et Mason (J. S.). – Dynamic constraint weighting in the context of articulatory parameter estimation. *In : Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 2535–2538. – Rhodes, Greece, September, 1997.

Relation entre pression sous glottique et intensité: étude des voyelles du francais

Véronique Lecuit[1*] et Didier Demolin[2]

[1]Laboratoire de Phonétique Expérimentale,
[2]Laboratoire de Phonologie,
Université Libre de Bruxelles,
av. F.-D. Roosevelt, CP110,
B-1050 Bruxelles– Belgique
Tél.: (32 2) 650 49 12 – Fax: (32 2) 650 20 07 – e-mail vlecuit@ulb.ac.be

ABSTRACT

This paper presents a study of the relationship between intensity and subglottic pressure in the production of french vowels during continuous speech. The study is based on direct measurement of subglottic pressure by tracheal puncture.

1. INTRODUCTION

Cette étude concerne la relation entre la pression sous-glottique (PSG) et l'intensité du signal de parole dans le cas des voyelles du français en parole continue. Une particularité de cette recherche provient du fait que nous disposons de la mesure directe de la PSG. Il nous a paru intéressant de nous pencher sur le problème du contrôle de l'intensité car c'est un problème complexe. En effet, selon Titze [Tit94], l'intensité peut être contrôlée au moyen de trois types de mécanismes: i) ajustement de la pression des poumons, ii) ajustement laryngeal, et iii) ajustement du conduit vocal. Plusieurs auteurs se sont interessés aux points i) et ii), notamment Ladefoged et McKinney [Lad63], Isshiki [Iss64], Strik et Boves [Str92], par mesure directe de PSG, Holmberg et al. [Hol88], McAllister et Sundberg [Mac96], par mesure indirecte de la PSG. En ce qui concerne les voyelles du français, Marchal et Carton [Mar79] présentent une étude sur la parole continue, en estimant la PSG à partir des valeurs de la pression oesophagienne corrigée en fonction des forces élastiques du poumon. Cependant, les données quantitatives restent rares, et tout nouvel apport en ce sens concourt à une meilleure compréhension des mécanismes impliqués. Ce papier présente les résultats obtenus sur un sujet masculin et un sujet féminin. Cette étude sur deux sujets ne constitue qu'un préliminaire à une étude plus vaste rassemblant 10 sujets francophones.

2. MÉTHODE

Deux sujets ont participé à cette étude. Un sujet francophone masculin de 43 ans, un sujet francophone féminin de 29 ans. Les deux sujets ne présentaient aucune pathologie du larynx et n'étaient pas entrainés pour la tâche. La pression sous glottique (PSG), la pression intra-orale (PIO), le débit d'air nasal, le débit d'air buccal et le signal de parole ont été enregistrés simultanément au moyen de la procédure suivante. Les enregistrements ont lieu à l'unité ORL de l'hôpital Erasme, Université de Bruxelles. Un tube flexible en caoutchouc (2mm diam. interne) est introduit à travers la cavité nasale jusqu'à l'oropharynx pour la mesure de PIO. Une aiguille (2mm diam. interne) est introduite dans la trachée après anesthésie locale (2% Xylocaine). Un tube du même type que le tube utilisé pour la PIO est connecté à l'aiguille. Le débit d'air buccal est mesuré avec un masque flexible en silicone. Le débit d'air nasal est mesuré au moyen d'une olive en silicone introduite dans la narine et reliée à un tube en plastique de 0.5 cm de diamètre. Tous les tubes et le masques sont reliés au polyphonomètre, station d'acquisition automatique de paramètres aérodynamiques de la parole, développée à l'institut de phonétique d'Aix (France) [Tes83]. Cette station possède un module de post-traitement des données qui nous a permis de calculer l'intensité (RMS) du signal de parole et la fréquence fondamentale (méthode du peigne). La tâche des sujets était de prononcer un corpus de 13 phrases du français, et ce à trois niveaux d'intensité différents: faible, normal, fort. Les sujets choisissant eux même pour chaque niveau l'intensité la plus confortable. Le choix des phrases est tel que l'on peut avoir un échantillon de voyelles dans des contextes divers. Nous avons ensuite relevé la mesure des différents paramètres au milieu de chaque voyelle. Après analyse, une partie des données correspondant au sujet masculin a du être éliminée, la mesure de PSG n'étant pas correcte (obstruction de l'aiguille). Toutes les plages d'intensité n'ont donc pas pu faire l'objet de mesure. Les données prises en considération sont 200 mesures réalisées sur l'ensemble des voyelles du sujet masculin, et 400 mesures sur l'ensemble des voyelles du sujet féminin.

3 RÉSULTATS

Le but de cette étude étant l'étude de la relation entre PSG et intensité pour les voyelles du français, nous avons tenu à présenter les résultats en séparant chaque voyelle pour chaque sujet (figures 1 à 4). Ceci nous permet de nous affranchir d'une éventuelle influence sur l'intensité due à la forme spécifique du conduit vocal liée à chaque voyelle. Afin de pouvoir comparer nos résultats avec les études existantes, nous avons calculé les équations de régression entre l'intensité et le logarithme de la PSG. Les coefficients des droites de régression (i ordonnée à l'origine, p pente) ainsi que le coefficient de corrélation (r) entre log PSG et intensité sont rassemblés dans une table (table 1). Pour chaque

paramètre, nous avons calculé la moyenne et l'écart type sur l'ensemble des voyelles. La dernière ligne de la table correspond au calcul de l'équation de régression sur l'ensemble du corpus pour chaque sujet.

Table 1: Coefficients de régression et coefficients de corrélation pour la relation entre log PSG et intensité.

	Sujet masculin			Sujet féminin		
	i	p	r	i	p	r
[a]	44.395	26.541	0.76	27.924	49.898	0.97
[œ]	39.466	32.652	0.81	30.055	44.490	0.94
[i]	35.631	27.860	0.79	24.586	45.832	0.89
[ɔ]	40.085	32.049	0.90	29.143	46.026	0.94
[y]	49.162	15.306	0.40	27.053	41.520	0.91
[e]	46.559	23.273	0.85	38.014	33.214	0.93
[ɛ]	44.149	26.083	0.84	31.970	42.304	0.97
[u]	44.597	22.134	0.66	25.797	43.412	0.96
[ø]	37.834	34.579	0.88	34.268	38.026	0.98
[o]	41.829	29.379	0.92	27.556	45.686	0.97
moy.	40.043	32.746	0.82	28.927	43.752	0.95
dev. st.	11.18	13.66	0.15	4.41	4.813	0.03
tout	41.79	27.83	0.67	31.725	40.79	0.86

Table 2: Incréments d'intensité (en dB) correspondant à un incrément de 2 hPa de PSG pour chaque voyelle et chaque sujet (F féminin, M masculin).

	[a]	[œ]	[i]	[ɔ]	[y]	[e]	[ɛ]	[u]	[ø]	[o]
F	15.0	13.4	13.8	13.8	12.5	10.0	12.7	13.1	11.4	13.7
M	8.0	9.8	8.4	9.6	4.6	7.1	7.8	6.7	10.4	8.8

Incréments d'intensité (en dB) correspondant à un incrément de 2 hPa de PSG

On observe un lien très étroit entre la PSG et l'intensité pour les deux sujets. Pour le sujet féminin, le coefficient de correlation associé à la droite de régression est en général largement supérieur à .90. En ce qui concerne le sujet masculin, on observe une moins bonne corrélation, ainsi que de très grandes variations d'une voyelle à l'autre de .40 pour [y] à .92 pour [o]. Ceci pourrait s'expliquer en partie par un manque de données dans une certaine gamme d'intensité, notament pour le [u]. Vu que notre méthode de mesure directe diffère fortement de la méthode indirecte utilisée par [Mar79] il est périlleux de vouloir comparer les résultats, cependant on observe le même type de relation entre PSG et intensité, tout en notant une moins grande dispersion des données relatives à chaque voyelle dans notre cas. La table suivante, (table 2), montre les incréments d'intensité (en dB) correspondant à un incrément de 2 hPa de PSG pour chaque voyelle et chaque sujet. Ces valeurs sont du même ordre de grandeur que celles publiées dans la littérature, 9.9 dB [Iss64], 13 dB [Hol88]. Il reste a vérifier si la différence entre le sujet masculin et le sujet féminin est liée à des différences inter-personnelles ou bien au sexe du sujet.

Conclusion

Au terme de cette étude, nous observons un lien très net entre la PSG et l'intensité. Il est donc nécessaire de s'interroger sur les facteurs contrôlant la la PSG. On peut émettre deux hypothèses: soit un contrôle au niveau des poumons, la variation de PSG provenant d'une variation du volume pulmonaire, soit un contrôle au niveau du larynx, provenant d'une variation de l'impédance glottique. Si l'on veut tester ces hypothèses il est très important, d'une part, de bien séparer les problèmes en préparant des protocoles adaptés à chaque hypothèse et d'autre part, de bien définir les paramètres pris en compte. On assiste trop souvent à la multiplication de paramètres dérivés à partir des mesures qui sont inutilisables, car trop liés au protocole expérimental utilisé. Si l'on considère par exemple la notion de seuil de phonation, la mesure de ce seuil dans les expériences MacAllister [Mac96] ne correspond pas a la définition stricte au sens où Titze l'entend [Tit94].

Bibliographie

[Hol88] Holmberg E.B., R.E. Hillman, J.S. Perkell. (1988), "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", J. Acoust. Soc. Am., Vol. 84, 2, pp. 511-529.

[Iss64] Isshiki N. (1964) "Regulatory mechanism of voice intensity variation", J. Speech Hearing Res., Vol. 7, pp. 17-29.

[Lad63] Ladefoged P. et N.P. McKinney. (1963) "Loudness, Sound Pressure, and Subglottal Pressure in Speech", J. Acoust. Soc. Am., Vol 35, 4. pp. 454-460.

[Mac96] Mc Allister A. et J. Sundberg. (1996), "Data on subglottal pressure and SPL at varied vocal loudness and pitch in 8 to 11-year-old children", TMH-QPSR, Vol. 4, pp. 27-35.

[Mar79] Marchal A. et F. Carton. (1979), "La pression sous-glottique: mesure et relation avec l'intensité et la fréquence fondamentale", Séminaire Larynx et Parole, ICP Grenoble, pp. 315-327.

[Str92] Strik H. et L. Boves. (1992), "Control of fundamental frequency, intensity, and voice quality in speech", Journal of Phonetics, Vol. 20, pp. 15-25.

[Tes83] Teston B. (1983), "Un systeme de mesure des paramètres aérodynamiques de la parole: le polyphonomètre modèle III", Travaux de L'Institut de Phonétique d'Aix. Vol. 9, pp. 373-383.

[Tit94] Titze I.R. (1994) Principles of Voice Production, Prentice Hall.

*Projet de la Communauté Française de Belgique, ARC 93/98-168.

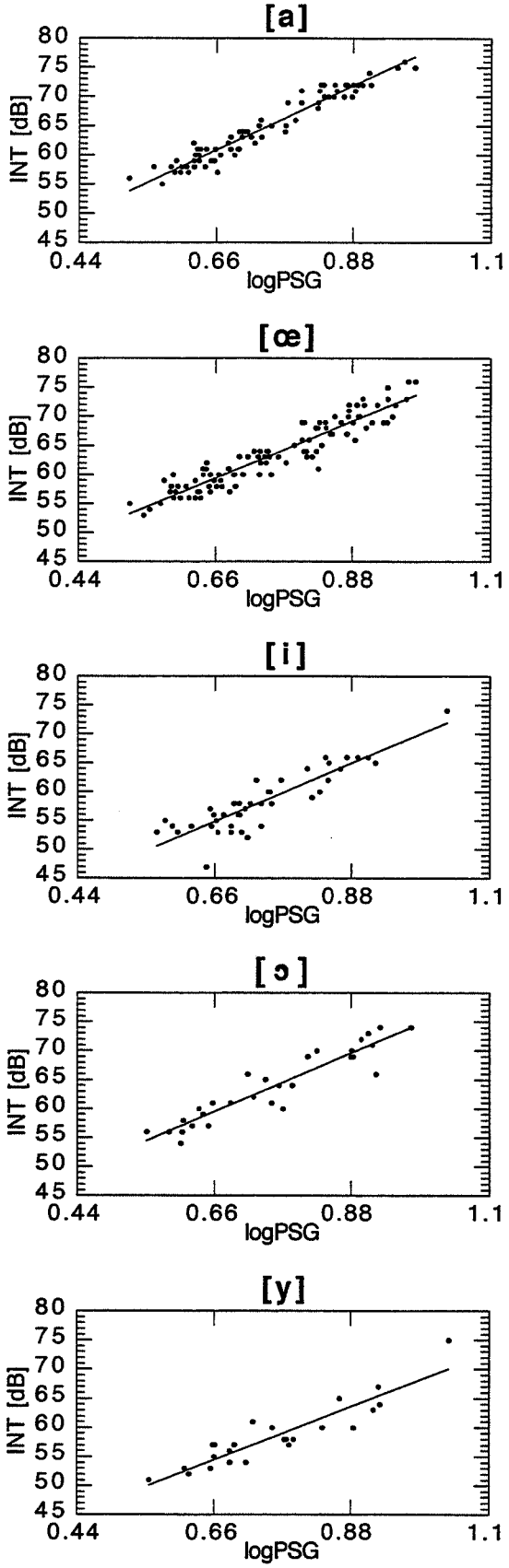


Figure1: Intensité vs log PSG sujet féminin

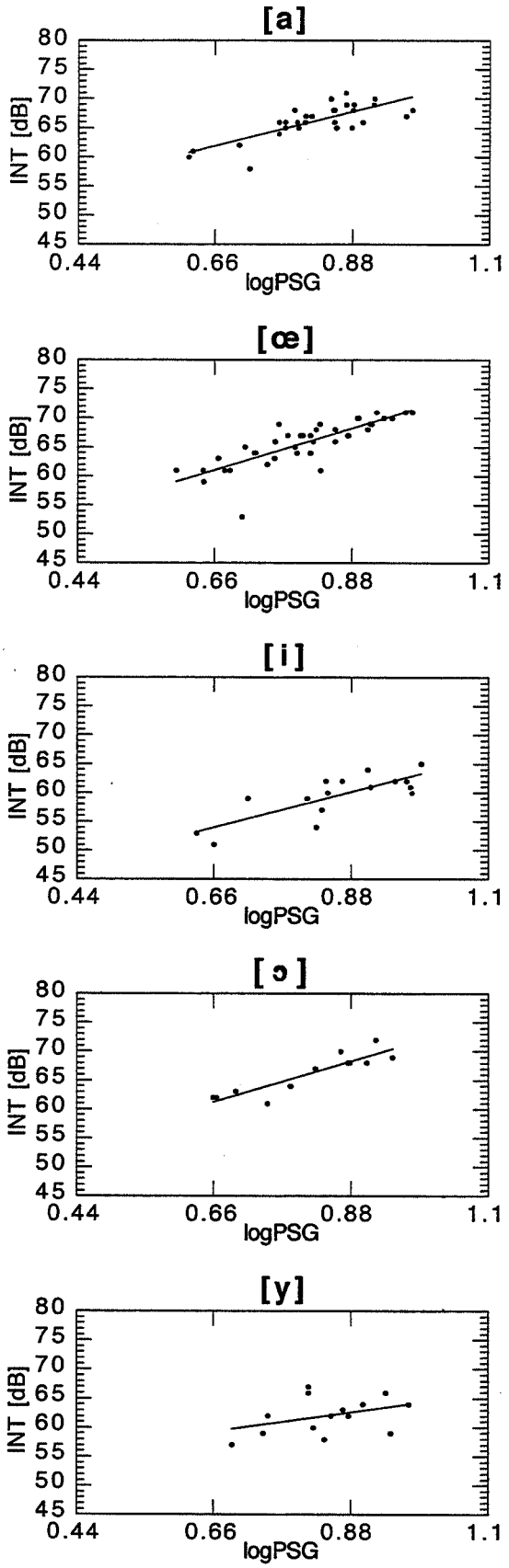


Figure2: Intensité vs log PSG sujet masculin

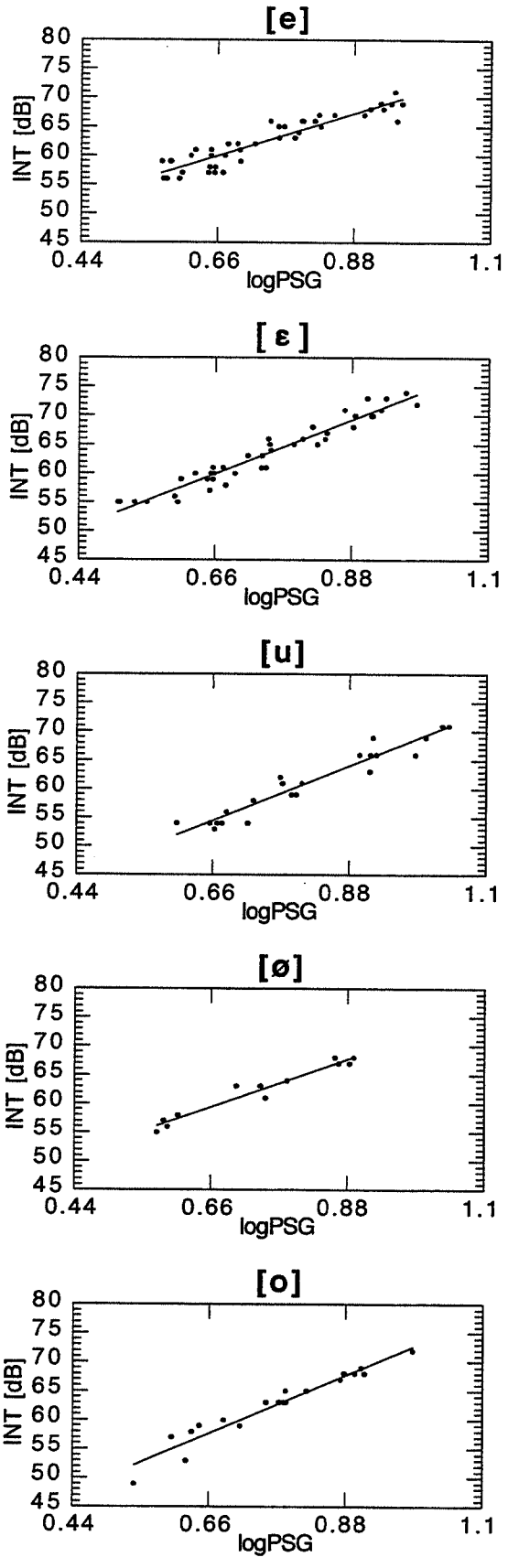


Figure3: Intensité vs log PSG sujet féminin

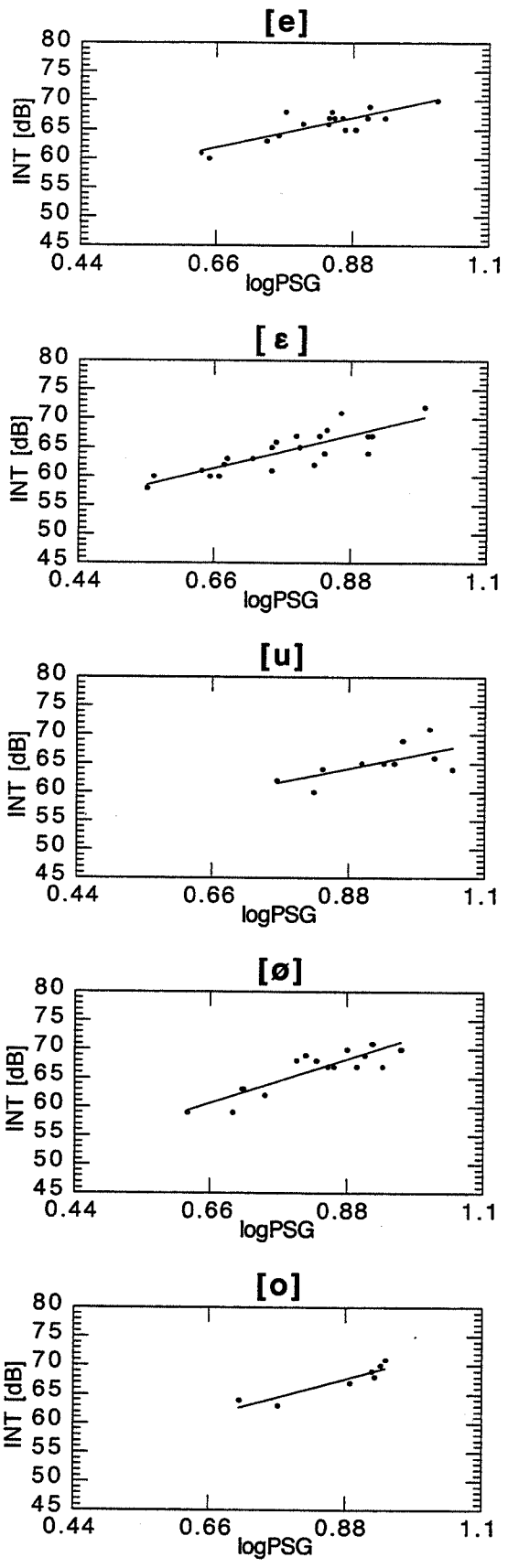


Figure4: Intensité vs log PSG sujet masculin

Frontière syllabique, accent focal contrastif et coarticulation linguale dans le groupe consonantique /kskl/ en position médiane de mot en français

Yohann Meynadier*, Alain Marchal° & Michel Pitermann*

* Laboratoire Parole et Langage (CNRS, ESA 6057)
Institut de Phonétique, Université de Provence
29 av. R. Schuman, F-13621 Aix-en-Provence cedex 1
Tél. : ++33 (0)4 42 95 36 24 – Fax : ++33 (0)4 42 59 50 96
e-mail : yohann.meynadier@lpl.univ-aix.fr – http://www.lpl.univ-aix.fr

° CNRS, DR 19, Caen, France

ABSTRACT

This paper explores the effects on the lingual articulation and coarticulation of the position of a contrastive focal accent at the left/right edges of the four-consonant cluster /kskl/ in medial position of the bisyllabic word "exclame" in French. Considering that an accent would affect more the segments belonging to the accented syllable than to the preceding or the following one, we suppose that the limit of the extent of his influences on the articulation of the different consonants and on the coarticulation within the cluster may correspond to the syllabic boundary. The results show that even if the position of a contrastive focal accent affects distinctively the coarticulation of the different sequences, especially concerning the coordination of the maximal constrictions, it's impossible by this way to localise exactly the syllable boundary.

1. INTRODUCTION

Nombre de travaux se sont posés la question de savoir dans quelle mesure l'observation de la coarticulation dans les groupes de consonnes permettrait de faire émerger l'organisation rythmique d'un énoncé, et donc par-là-même les différentes unités de regroupement syllabique, lexicale ou prosodique. L'hypothèse est que plus le chevauchement est important et moins il est variable, plus la cohésion de la séquence est forte, donc plus elle a de chance d'appartenir à une même unité d'organisation.

S'agissant de /kl/, Hardcastle [Har85] a montré que lorsque chaque consonne se trouvait de part et d'autre d'une frontière majeure de fin de regroupement prosodique (et syntaxique), une tendance à l'augmentation de l'intervalle interconsonantique pouvait être observée en parole lente. Concernant la frontière de mot, les travaux de Byrd [Byr96] ont montré que les séquences /dg/, /gd/, /ks/ et /sk/ étaient caractérisées par un chevauchement moins important et un timing articulatoire intrinsèque moins variable en début de mot qu'en fin ou à cheval entre deux unités lexicales. Il apparaît donc dans une certaine mesure que des régularités dans la coordination articulatoire des séquences consonantiques laissent supposer l'existence d'un marquage articulatoire de frontières prosodiques (d'un point de vue rythmique).

Dans cette étude, nous nous sommes concentrés sur un niveau plus bas de regroupement syntagmatique : celui de la syllabe (en dehors de toutes influences pouvant être induites par une frontière lexicale). Or, si les chercheurs partagent assez largement l'idée que la syllabe a une valeur phonologique et fonctionnelle, sa réalité phonétique (acoustique et articulatoire) reste encore hasardeuse [Ble94]. Afin d'essayer de dégager des corrélats articulatoires à la frontière syllabique, nous avons choisi de nous attacher à l'organisation articulatoire interne au groupe de quatre consonnes linguales /kskl/ en position médiane de mot en français.

2. HYPOTHESE

La méthode expérimentale utilisée ici repose sur l'examen des effets d'un accent focal contrastif (AFC) sur la coarticulation linguale dans le groupe /kskl/. L'accent étant supposé d'une part renforcer l'articulation [Jon95] et d'autre part réduire les phénomènes de coarticulation [Jon93], il est attendu que son influence soit distinctement plus grande pour les consonnes du groupe appartenant à la syllabe accentuée plutôt que pour celles rattachées à la syllabe précédente ou suivante. En cela, la limite de l'influence de l'AFC dans le groupe selon sa position par rapport à celui-ci (à sa droite ou à sa gauche) pourrait correspondre à la frontière syllabique.

3. METHODE

3.1. Sujets et corpus

Le corpus est composé du groupe quadriconsonantique /kskl/ situé au milieu d'un mot bisyllabique et ayant les voyelles ouvertes /ɛ_a/ pour contexte vocalique (« exclame »). Ce mot est inséré au milieu d'une phrase porteuse lue à débit normal par deux locuteurs français natifs (B et Y). Les locuteurs étaient intuitivement contraints d'accentuer alternativement trois syllabes différentes de la phrase en fonction de la question posée par l'expérimentateur simulant l'incompréhension de la syllabe cible à accentuer.

Ainsi, le groupe a été analysé selon trois positions accentuelles : 1) /kskl/ en position inaccentuée (A0) : « J'ai *dit* : "Pam exclame même sa peine" », répondant à « Tu as *pensé* : "Pam exclame même sa peine" » (10 répétitions) ; 2) /kskl/ accentué à gauche (A1) : « J'ai dit : "Pam *ex*clame même sa peine" », répondant à « Tu as dit : "Pam *dé*clame même sa peine ?" » (20 répétitions) ; 3) /kskl/ accentué à droite (A2) : « J'ai dit : "Pam ex*clame* même sa peine" », répondant à « Tu as dit : "Pam ex*prime* même

sa peine ?" » (20 répétitions).

Nous avons vérifié que les accents se trouvaient bien à la position attendue et qu'ils se distinguaient nettement de la condition non accentuée. Pour cela, une analyse de la variance réalisée sur les valeurs de fréquence fondamentale de V_1 et V_2, mesurées au milieu de la voyelle, a montré que pour les deux locuteurs ces voyelles avaient respectivement une F0 très significativement plus haute en contexte accentué qu'inaccentué ($p < .0001$).

3.2. Acquisition et analyse EPG

Les contacts de la langue avec le palais dur ont été enregistrés à une fréquence d'échantillonnage de 200 Hz au moyen d'un appareillage électropalatographique muni de 62 électrodes (Reading EPG System). Ce signal EPG permet de rendre compte de la chronologie et de la localisation des constrictions linguo-palatales.

La méthode d'analyse et de réduction des données EPG utilisée ici est empruntée à Byrd [Byr95]. Elle permet de définir une zone d'articulation EPG spécifique pour chaque consonne linguale. Ainsi, nous avons défini pour chaque locuteur une zone postérieure (relative à /k/) et une zone antérieure (relative à /s/ et /l/) hermétiques l'une à l'autre. En calculant simultanément dans chaque zone et à chaque trame, le pourcentage d'électrodes contactées par rapport au nombre total d'électrodes de la zone, il est possible de rendre compte de l'évolution spatio-temporelle des différentes constrictions linguo-palatales impliquées dans le groupe consonantique /kskl/ (figure 1).

Les items présentant une fusion ou une élision complète de gestes consonantiques ont été retirés de l'analyse.

3.3. Mesures temporelles

Les durées de différentes phases articulatoires ont été calculées à partir de quatre événements articulatoires considérés comme pertinents dans l'articulation d'une consonne : le début et la fin de la constriction linguo-palatale (ou consonne) représentés respectivement par la première et la dernière image EPG montrant au moins un contact dans sa zone d'articulation, et, le début et la fin de la tenue articulatoire d'une consonne correspondant respectivement à la première et la dernière image EPG montrant le maximum de contacts dans sa zone d'articulation.

Ainsi, la durée globale d'une consonne ou d'une séquence biconsonantique *(DurSQ)* est définie respectivement par le temps qui sépare le début d'une consonne de la fin de cette même consonne ou de la consonne suivante. La durée de l'attaque *(DurON)*, de la tenue *(DurTN)* et du relâchement *(DurOF)* de chaque consonne correspondent respectivement à l'écart temporel entre le début de la consonne et de sa tenue, entre le début et la fin de la tenue et entre la fin de la tenue et celle de la consonne. Trois autres mesures temporelles sont relatives à la coordination articulatoire des constrictions linguo-palatales. La durée du chevauchement interconsonantique est déterminée par l'écart temporel entre la fin d'une consonne et le début de la suivante (*Chev* concerne le chevauchement dans les séquences biconsonantiques ; *Chev1*, celui d'une consonne avec celle qui la précède ; *Chev2*, avec la suivante). Il est apparu que tous les items sauf un (B et Y confondus) présentaient un chevauchement articulatoire. La durée de la partie non chevauchée d'une consonne est définie par la durée de cette consonne moins celui du chevauchement avec la consonne précédente *(NoChev1)* ou suivante *(NoChev2)*. L'intervalle d'inter-tenue d'une séquence biconsonantique *(InterTN)* correspond à la durée entre la fin de la tenue de la première consonne et le début de celle de la deuxième. Cet intervalle est négatif si les deux tenues se chevauchent.

4. RESULTATS

Les données ont été traitées statistiquement au moyen d'analyses de la variance post-hoc (PLSD de Fischer à un facteur, Table 1) et de corrélations (coefficient R^2 de régression linéaire entre deux facteurs, Table 2).

4.1. Influences de l'AFC sur l'articulation et la coarticulation des consonnes linguales

La Table 1 montre assez nettement pour B et Y que la présence d'un accent aux bornes des séquences /ks/, /sk/ et /kl/ augmente significativement leur durée, du fait de l'allongement des consonnes (excepté /l/). Cependant, l'augmentation de la durée consonantique n'est pas uniforme. Seule la durée de la phase de relâchement de la consonne est significativement et systématiquement affectée par la présence d'un AFC, alors que celle de l'attaque et de la tenue ne semblent pas en être dépendantes.

La Table 1 fait également apparaître qu'un AFC a pour effet d'augmenter la durée de la partie non chevauchée des consonnes (sauf $/k/_2$ et /l/ pour Y) et plus étonnement de la partie chevauchée des séquences (sauf /kl/). Or, la durée de ces deux phases semble, dans la plupart des cas, être en bonne partie corrélée à la durée consonantique. Il ressortirait donc que l'allongement de ces phases n'est pas directement dû à la présence de l'AFC, mais simplement à l'allongement des consonnes en contexte accentué. De plus, en terme de proportion de la durée consonantique, ces phases ne sont pas corrélées à la durée des consonnes (Table 2). Il est donc possible que l'allongement des consonnes, imposé directement par la présence d'un AFC, se répercute proportionnellement sur les phases de chevauchement et de non-chevauchement des gestes.

Concernant l'intervalle inter-tenue des séquences, la Table 2 montre que sa durée, comme sa proportion de la durée de la séquence, n'est pas corrélée à la durée globale des séquences. Or, pour /sk/ et /kl/, cette phase interconsonantique est significativement allongée par la présence d'un AFC (Table 1). Il apparaît donc que la présence d'un AFC influence directement la coordination des phases de tenue articulatoire des consonnes.

4.2. Empan dans le groupe de l'influence de l'AFC selon sa position

L'examen de la limite dans le groupe de l'influence d'un

AFC selon sa position montre une différence importante entre les deux locuteurs. Globalement, pour Y seule la présence d'un accent, quelle que soit sa position, a une influence sur l'articulation du groupe ; tandis que B montre plus de différences entre les trois conditions accentuelles. Les résultats sont résumés dans la Table 1.

L'influence dominante sur le groupe d'un AFC initial de mot (A1) par rapport à A2 semble pour B se limiter à la consonne initiale du groupe ($/k/_1$), alors que pour Y celle-ci remonte jusqu'à $/k/_2$. Malgré tout, il apparaît pour les deux locuteurs que son influence distinctive de la condition inaccentuée s'applique sur toute la longueur du groupe. Concernant l'AFC final de mot (A2), son influence dominante par rapport à A1 sur le groupe semble pour B pouvoir remonter jusqu'à /s/, tandis que pour Y elle se limite à /l/. Toutefois, son influence distinctive de la condition inaccentuée s'établit sur tout le groupe pour B et jusqu'à /s/ pour Y.

Plus en détails, à travers les différents paramètres articulatoires il ressort assez nettement pour les deux locuteurs que /l/ est spécifiquement influencé par A2. Bien que la présence d'un AFC final réduit significativement sa durée, cet effet inverse de l'accent sur /l/, par rapport aux autres consonnes, peut s'expliquer par un abaissement de la mandibule inférieure probablement plus ample, tôt et/ou rapide en contexte accentué (surtout A2) du fait de la présence de la voyelle ouverte suivante /a/, ce qui affecte considérablement la durée du geste apical [Jon95]. Cela est particulièrement net pour Y, alors que pour B le fait que cette réduction n'apparaisse qu'en contexte A1 est trompeur. En effet, dans ce cas, /l/ subit deux influences contradictoires : d'une part, une réduction de sa durée du fait de la contrainte exercée par l'articulation de /a/ en contexte accentué, et d'autre part, une augmentation de la durée des consonnes linguo-palatales en contexte accentué. Alors que pour Y la contrainte articulatoire exercée par la voyelle ouverte accentuée prend le dessus, pour B il y a un conflit entre ces deux influences en contexte A2, ce qui aboutit à un compromis articulatoire qui neutralise l'effet distinctif de A2 sur la durée globale de /l/ (en moyenne, A0 : 97 ms ; A1 : 78 ms ; A2 : 93 ms).

Contrairement à Y, l'influence dominante de A2 sur la partie finale du groupe allant de /s/ à /l/ ressort assez bien pour B. En effet, les séquences /sk/, /kl/ et /skl/, ne contenant pas la consonne initiale $/k/_1$, voient leur durée globale et leur intervalle d'inter-tenue augmenter sous l'effet exclusif de A2.

Néanmoins, pour les deux locuteurs, l'AFC n'influence en rien l'écart temporel entre les tenues articulatoires des consonnes impliquées dans /ks/.

5. DISCUSSION

Tout d'abord, il ressort que l'attaque et la tenue de la constriction linguo-palatale d'une consonne, impliquée dans un groupe consonantique intralexical, apparaissent être des phases articulatoires relativement stables en durée entre une production en contexte inaccentué et accentué (AFC). Cette stabilité avait déjà été notée dans une étude antérieure [Mar96] au regard du style de production (logatome/mot isolé *vs* phrase). Par contre, les phases d'enchaînement interconsonantique semblent particulièrement sensibles aux contextes accentuel et stylistique. Ici, cela paraît être principalement le fait de la grande dépendance au contexte accentuel de la durée de la phase de relâchement d'une consonne. La coordination articulatoire entre deux constrictions linguo-palatales reposerait donc en bonne partie sur la durée de cette phase.

Il apparaît également qu'un AFC n'impose pas à proprement parler une réduction de la coproduction de gestes de constriction linguo-palatale. En effet, l'AFC semble n'influencer qu'indirectement la durée du chevauchement articulatoire entre deux consonnes contiguës : la durée de celui-ci dépendant principalement de celle des consonnes. Or, privilégiant l'allongement des consonnes, il résulte qu'un AFC favorise un allongement du chevauchement interconsonantique. Ainsi, nos résultats semblent aller à l'encontre de l'intuition selon laquelle un accent réduirait globalement les phénomènes de coproduction de gestes. Par contre, l'AFC, en imposant directement l'augmentation de l'écart temporel entre deux constrictions maximales, renforce l'émergence articulatoire (et probablement perceptive) des tenues et donc des parties stables des gestes linguo-palataux. Ainsi, c'est principalement à ce niveau de la coordination articulatoire que la coarticulation entre deux gestes consonantiques linguo-palataux adjacents est réduite.

Enfin, si en règle générale une influence distinctive de l'accent en fonction de sa position est visible, il ressort malgré tout que, quelle que soit sa position, la limite de l'influence d'un AFC sur l'articulation des consonnes et sur la coarticulation des séquences ne peut être nettement établie. En effet, l'AFC marque non seulement la syllabe sur laquelle il porte, mais également la syllabe précédente ou suivante. Il paraît donc délicat de dégager une rupture dans l'influence décroissante d'un AFC selon que les consonnes se situent de plus en plus loin de la syllabe accentuée. L'un des problèmes majeurs repose sur la grande différence des contraintes articulatoires et/ou aérodynamiques entre les séquences biconsonantiques composant ce groupe. Le comportement face à l'AFC de l'intervalle inter-tenue de /ks/ et de /sk/ en est exemplaire. En effet, à l'inverse de /ks/, dans /sk/ une exigence aérodynamique essentielle contraint la coordination articulatoire entre la tenue de /s/ et de /k/ : le début de la tenue de /k/ ne doit pas intervenir trop tôt par rapport à la fin de celle de /s/ sous peine d'en interrompre brutalement le bruit de friction. L'existence de cette contrainte importante impose donc à l'AFC de renforcer cet écart des tenues dans /sk/. Par contre, l'absence de cet impératif aérodynamique dans /ks/ explique que, dans ce cas, l'écart temporel entre la tenue de /k/ et celle de /s/ n'est pas spécifiquement renforcé par l'AFC (Table 1).

6. CONCLUSION

Les résultats de cette étude montrent 1) qu'un accent focal contrastif affecte essentiellement le relâchement de la

constriction linguo-palatale d'une consonne, alors que la durée des phases d'attaque et de tenue restent constantes ; 2) qu'un accent focal contrastif n'entraîne pas de réduction de la coproduction des gestes, mais que par contre il impose directement un écart temporel plus important entre les tenues articulatoires (constrictions maximales) de deux consonnes linguo-palatales contiguës ; 3) qu'un accent focal contrastif influence la coordination articulatoire entre deux consonnes linguo-palatales en contact plus en fonction des contraintes articulatoires intrinsèques à la séquence que des contraintes rythmiques liées au marquage de la frontière syllabique. Ainsi, du fait de l'interférence des contraintes articulatoires, la manipulation d'un accent focal contrastif ne semble donc pas pouvoir permettre de localiser efficacement la frontière syllabique intervenant dans le groupe consonantique /kskl/ vis-à-vis de sa coordination gestuelle interne.

BIBLIOGRAPHIE

[Ble94] Blevins J. (1994) "The syllable in the phonological theory", in J. Goldsmith (ed.), The Handbook of Phonological Theory, pp. 206-244. Oxford : Blackwell Publisher.

[Byr95] Byrd D., Flemming E., Mueller C. & Cheng Tan C. (1995) "Using regions and indices in EPG data reduction", Journal of Speech and Hearing Research, Vol. 38, pp. 821-827.

[Byr96] Byrd D. (1996) "Influences on articulatory timing in consonant sequences", Journal of Phonetics, Vol. 24(2), pp. 209-244.

[Har85] Hardcastle W. (1985) "Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences", Speech Communication, Vol. 4, pp. 247-263.

[Jon95] Jong de K. (1995) "The supraglottal articulation of prominence in English : linguistic stress as localized hyperarticulation", Journal of the Acoustical Society of America, Vol. 97(1), pp. 491-504.

[Jon93] Jong de K., Beckman M. & Edwards J. (1993) "The interplay between prosodic structure and coarticulation", Language and Speech, Vol. 36 (2-3), pp.197-212.

[Mar96] Marchal A. & Meynadier Y. (1996) « Coarticulation in /kl/ sequences in French : A multisensor investigation of the timing of lingual gestures », Proceedings of the 6th Australian International Conference on Speech Science and Technology, pp. 435-438. Adelaide.

REMERCIEMENTS

Cette étude a été réalisée dans le cadre d'un contrat de recherche financé par le Conseil Régional Provence-Alpes-Côte d'Azur. Nous tenons également à remercier tout particulièrement B. Lagrue (LPL), B. Galindo (Société SQ Lab) et R. Espesser (LPL)

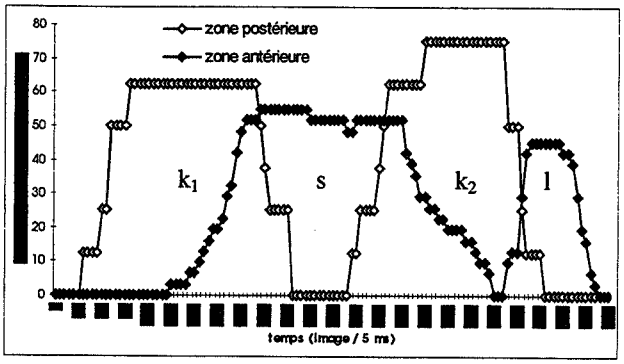


Figure 1 : Courbes d'évolution spatio-temporelle des constrictions linguo-palatales impliquées dans /kskl/

Table 1 : Effet de l'AFC en fonction de sa position sur la durée des phases pour chaque locuteur : seul est indiqué le type d'accent qui a un effet significatif, d'augmentation ou de réduction (en gras), sur la durée des phases par rapport à A0 (p.< .05). "1" = A1 ; "2" = A2 ; "3" = A1 et A2 indifférenciés ; ">" = l'AFC à gauche a un effet significativement plus marqué ; "*" = l'AFC indiqué se distingue significativement de A0 mais pas de l'autre condition accentuée, elle-même indifférenciée de A0 ; "-" = absence de différence significative entre A0, A1 et A2

Indice	*Loc*	*k1*	*s*	*k2*	*l*	*ks*	*sk*	*kl*	*ksk*	*skl*
NoChev1	B		3	1*	-					
	Y		3	3	2					
NoChev2	B	1	2*	3						
	Y	-	3	3						
DurON	B	-	-	-	2					
	Y	-	-	-	-					
DurTN	B	-	-	2*	-					
	Y	-	3	-	-					
DurOF	B	3	2>1	2	2>1					
	Y	1*	3	1*	2*					
DurSQ	B	3	2>1	2	1	3	2	2	3	2
	Y	1*	3	1*	2>1	3	3	-	3	3
Chev	B					2	3	1*		
	Y					-	3	3		
InterTN	B					-	2	2>1		
	Y					-	3	3		

Table 2 : Corrélations (R^2) entre la durée des consonnes ou des séquences et la durée des phases en ms et en proportion de la durée des consonnes ou des séquences pour chaque locuteur (en italique : corrélations négatives)

*DurSQ**	*Loc*	*k1*	*s*	*k2*	*l*	*ks*	*sk*	*kl*
Chev1	B		0,57	0,55	0,63			
	Y		0,65	0,73	0,17			
NoChev1	B		0,72	0,003	0,11			
	Y		0,34	0,001	0,67			
NoChev1 %	B		0,20	*0,11*	*0,21*			
	Y		*0,04*	*0,20*	*0,001*			
Chev2	B	0,39	0,74	0,01				
	Y	0,44	0,46	0,03				
NoChev2	B	0,34	0,62	0,54				
	Y	0,04	0,84	0,62				
NoChev2 %	B	0,13	*0,20*	0,09				
	Y	*0,13*	0,08	0,05				
Chev	B					0,34	0,41	*0,004*
	Y					0,13	0,38	*0,0004*
Chev %	B					*0,22*	0,06	*0,14*
	Y					*0,002*	*0,01*	*0,13*
InterTN	B					*0,03*	0,05	0,30
	Y					0,02	0,11	0,005
InterTN %	B					*0,01*	0,03	0,23
	Y					*0,05*	0,05	*0,01*

La relation entre les niveaux articulatoire et acoustique peut être décrite par des équations différentielles couplées

J. Schoentgen[1], S. Ciocea[2]

Laboratoire de Phonétique Expérimentale, Université Libre de Bruxelles, C.P. 110
50, Av. F.-D. Roosevelt, B-1050 Bruxelles
Tél: +32 2 650 2010 - Fax + 32 2 650 2007, E-mail: jschoent@ulb.ac.be

ABSTRACT

The objective is to present a formalism which offers a common framework for several articulatory models and concepts such as targets, gestures and the quantal principle of speech production. The formalism is based on coupled differential equations which relate the vocal tract shape to its eigenfrequencies. The shape of the vocal tract is described either directly by means of an area function model, or indirectly by means of an articulatory model. The synergetic relation between phonetic gestures and targets and the quantal principle of speech production is also discussed.

1. INTRODUCTION

L'objectif est de présenter un formalisme mathématique qui réunit des modèles articulatoires et acoustiques, ainsi que des concepts tels que la cible, le geste, et le principe quantal de la production de la parole.

Ce cadre formel est une généralisation d'un développement antérieur grâce auquel nous avons formulé une solution analytique au problème de l'inversion acoustico-géométrique [Sch97a]. L'obtention de cette solution démontre la pertinence de la formulation mathématique [Cio97b], [Pit96a], [Sch97b].

Le formalisme est fondé sur des équations différentielles couplées reliant la forme du conduit vocal à ses fréquences propres. La forme est décrite soit directement par la fonction d'aire, soit indirectement par un modèle articulatoire. La fonction d'aire est la section droite du conduit vocal en fonction de la distance à la glotte. Les avantages de cette formulation sont les suivants. Premièrement, la formulation différentielle est locale et explicite tandis que la formulation algébrique conventionnelle est globale et implicite. Une fonction implicite est caractérisée par la présence des variables dépendantes et indépendantes du même côté du signe d'égalité. Deuxièmement, les équations différentielles mettent en évidence la cinétique de la forme du conduit vocal. Elles facilitent ainsi la discussion des concepts de cibles et de gestes. Troisièmement, ce formalisme est à la base d'une solution analytique du problème de l'inversion acoustico-géométrique, c.-à-d. d'une méthode mathématique qui infère la fonction d'aire à partir des fréquences des formants [Cio97a]. Quatrièmement, le formalisme mathématique permet la représentation précise d'une série de concepts propres à différents modèles de production de la parole, mais il se prête également à la simulation numérique des étages acoustiques ou articulatoires.

Notre article privilégie la discussion des concepts. Le texte est organisé comme suit. La section 2. développe la relation différentielle entre un modèle de la fonction d'aire et les fréquences propres. La section 3. étend ces relations aux modèles paramétriques de la fonction d'aire ou des articulateurs. La section 4. discute les implications de ces relations pour les concepts de cibles articulatoires ou acoustiques. La section 5. présente un modèle gestuel. La discussion, finalement, concerne les liens synergiques entre cibles et gestes articulatoires d'une part et le principe quantal de la production de la parole d'autre part.

2. FORMULATION DIFFERENTIELLE DE LA RELATION ENTRE FONCTION D'AIRE ET FREQUENCES PROPRES

La relation est fondée sur l'équation de Webster qui décrit la propagation d'une onde acoustique plane dans un tube sans pertes et de forme quelconque. Des solutions analytiques existent pour plusieurs formes de tube, dont le cylindre [Bon83]. En ce cas, les modes propres sont décrits par la relation matricielle (1). Les modes propres sont les solutions de l'équation de Webster pour lesquelles l'onde oscille avec la même fréquence et la même phase en tout point du tube. La connaissance des modes propres est essentielle parce que tout champ acoustique décrit par l'équation de Webster est une somme pondérée des modes propres. Cette propriété mathématique explique l'identité entre fréquences propres et fréquences des formants observées.

[1] Fonds National de la Recherche Scientifique, Belgique
[2] Action de Recherche Concertée, Communauté Française de Belgique

La relation entre les niveaux articulatoire et acoustique peut être décrite par des équations différentielles couplées

Les pressions et vitesses acoustiques à l'entrée et à la sortie du cylindre sont notées p_e, v_e, p_s, et v_s. Les symboles S_j et l_j sont la section droite et la longueur du cylindre; ω la fréquence propre, i la racine de -1; c la vitesse du son et ρ la densité de l'air.

$$\begin{pmatrix} p_e \\ u_e \end{pmatrix} = \begin{pmatrix} \cos\frac{\omega l_j}{c} & i\frac{\rho c}{S_j}\sin\frac{\omega l_j}{c} \\ i\frac{S_j}{\rho c}\sin\frac{\omega l_j}{c} & \cos\frac{\omega l_j}{c} \end{pmatrix} \begin{pmatrix} p_s \\ u_s \end{pmatrix} \quad (1)$$

La forme du conduit vocal peut être approchée à l'aide d'une concaténation de N tubes cylindriques. La relation (2) entre les pressions et vitesses aux lèvres et à la glotte est alors obtenue en multipliant N matrices de transfert (1).

$$\begin{pmatrix} p_{glot} \\ u_{glot} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} p_{lev} \\ u_{lev} \end{pmatrix}. \quad (2)$$

A la glotte et aux lèvres, le tube vocal est en contact avec l'extérieur. Par conséquent, en ces endroits des conditions aux bords doivent être imposées. Les conditions les plus simples sont $u_{glot} = 0$ (paroi infiniment rigide) et $p_{lèv} = 0$ (milieu infiniment souple). La matrice (2) se réduit alors à l'équation non-linéaire suivante [Bon83].

$$D(S_j, l_j, \omega) = 0. \quad (3)$$

L'équation (3) réunit la forme du conduit et le modèle de la propagation acoustique. La forme de l'équation (3) reste d'ailleurs inchangée pour d'autres modèles de la fonction d'aire et de la propagation acoustique [Jos95], [Cio96].

L'équation (3) permet en effet de calculer les fréquences propres ω_k du conduit en fonction de la forme imposée. Pendant la production de parole continue, la forme du conduit est évolutive. En supposant valable l'équation de Webster sous ces conditions, nous pouvons rendre explicite l'évolution temporelle de la forme du conduit en dérivant l'équation (3) par rapport au temps. Le résultat est le système d'équations couplées (4).

$$\frac{dD}{dt} = \sum_{j=1}^{N} \frac{\partial D}{\partial S_j}\frac{dS_j}{dt} + \frac{\partial D}{\partial \omega_k}\frac{d\omega_k}{dt} = 0, \quad k = 1...M \quad (4)$$

M est le nombre de fréquences propres (typiquement 3) et N le nombre de cylindres concaténés. Afin de simplifier la notation, les longueurs des cylindres sont supposées fixes. Les dérivées partielles sont calculées soit analytiquement soit numériquement à l'aide de l'équation (3). Le système d'équations différentielles linéaires (4) peut être résolu itérativement par des méthodes conventionnelles. Les conditions initiales sont choisies de manière à satisfaire la condition D=0.

3. FORMULATION DIFFERENTIELLE DE LA RELATION ENTRE MODELES ARTICULATOIRES ET FREQUENCES PROPRES

Souvent la forme du conduit vocal n'est pas décrite directement à l'aide des sections droites mais à l'aide d'un modèle paramétrique de la fonction d'aire ou à l'aide d'un modèle articulatoire. Un modèle paramétrique de la fonction d'aire est la silhouette stylisée du conduit vocal. Un modèle articulatoire est une représentation simplifiée de la forme et des postures des articulateurs dans le plan sagittal. La raison d'être de ces modèles est la possibilité qu'ils offrent de contrôler la forme du conduit vocal à l'aide d'un faible nombre de paramètres articulatoirement ou phonétiquement pertinents. Sauf exception, ces modèles sont purement géométriques et ne font pas explicitement référence aux forces musculaires ou aux masses et élasticités des articulateurs. Ci-après, nous désignons par θ_i les paramètres de contrôle des modèles et par δ_j les distances inter-sagittales.

Formellement, le lien entre T paramètres articulatoires θ et N distances inter-sagittales δ s'écrit de la manière suivante.

$$\frac{d\delta_j}{dt} = \sum_{i=1}^{T} \frac{\partial \delta_j}{\partial \theta_i}\frac{d\theta_i}{dt}, \quad j = 1...N \quad (5)$$

Mathématiquement parlant, l'expression (5) est le résultat de la multiplication de la matrice $\left[\frac{\partial \delta_j}{\partial \theta_i}\right]$ avec un vecteur qui contient les dérivées temporelles des paramètres articulatoires. La matrice comprend toutes les combinaisons possibles des dérivées des distances par rapport aux paramètres. Dans le cas d'un modèle articulatoire, elle n'est pas nécessairement carrée car le nombre de paramètres θ n'est pas toujours égal au nombre de distances δ ou sections droites S. Si la matrice était carrée, c.-à-d. si N = T, le passage des paramètres articulatoires aux sections droites serait un changement de variables réversible (et l'inversion acoustico-articulatoire en serait grandement simplifiée). Le lien heuristique entre distances sagittales et section droites est exprimé à l'aide de la relation (6) [Hei65]

$$S_j = \alpha \delta_j^{\beta}. \quad (6)$$

En substituant les expressions (5) et (6) dans les équations différentielles couplées (4), un nouveau système d'équations (7) est obtenu qui relie paramètres articulatoires θ_i et fréquences propres ω_k.

$$\sum_{j=1}^{N} \frac{\partial D}{\partial S_j}\alpha\beta\delta_j^{\beta-1}\left(\sum_{i=1}^{T}\frac{\partial \delta_j}{\partial \theta_i}\frac{d\theta_i}{dt}\right) + \frac{\partial D}{\partial \omega_k}\frac{d\omega_k}{dt} = 0,$$
$$k = 1...M. \quad (7)$$

Les dérivées temporelles des paramètres articulatoires sont co-déterminés par le module de commande du modèle de production choisi. Hormis les dérivées $\frac{d\theta_i}{dt}$, toutes les dérivées partielles peuvent être obtenues

numériquement ou analytiquement, soit à l'aide du modèle articulatoire, soit à l'aide de l'équation D=0. L'approximation numérique des dérivées partielles se fait en divisant les variations des variables au numérateur provoquées par une légère variation des variables au dénominateur. Finalement, avant de pouvoir utiliser le système (7) à des fins de calculs, des conditions initiales doivent être imposées, qui sont une solution de l'équation D = 0.

Nous proposons d'examiner dans quelle mésure les concepts de cible et geste phonétique ainsi que le principe quantal peuvent être dérivés à partir de la relation (7).

4. CIBLES ACOUSTIQUES ET ARTICULATOIRES

La cible articulatoire est une configuration quasi-stationnaire des organes de parole que le locuteur tenterait d'atteindre lors de la production d'un segment phonétique. La cible acoustique est la transposition de cette définition au niveau des fréquences des formants [Lin63].

Mathématiquement parlant, l'atteinte d'une cible acoustique implique $\frac{d\omega_k}{dt} = 0,\ k = 1...M$. Cela signifie que dans le système (7), l'expression à gauche du signe d'addition s'annule. Des solutions particulières de ce nouveau système d'équations sont les suivantes.

1) $\frac{d\theta_i}{dt} = 0,\ i = 1...T.$ Cela signifie qu'une cible articulatoire a été atteinte et que les postures des articulateurs sont momentanément figées.

2) $\frac{\partial \delta_j}{\partial \theta_i} = 0,\ i = 1...T, j = 1...N.$ Cette situation se présente lorsque les articulateurs bougent sans changer la forme du conduit. La langue monte, par exemple, tandis que la mâchoire descend. Ou encore, les lèvres s'écartent pendant que la mâchoire monte. Ces manoeuvres sont évidemment inappropriées pour réaliser des oppositions phonétiques.

3) $\frac{\partial D}{\partial S_j} = 0,$ j = 1...N. Lorsque cette condition est remplie, un lien hyposensible existe entre forme et fréquences. Cela signifie que des changements au voisinage de cette forme n'entraînent pas des changements des fréquences propres. La théorie qui accorde un rôle phonétique particulier à ces formes du conduit est connue sous le nom de théorie quantale de la production de la parole [Ste89]. Le système (7) suggère que cette condition est équivalente à $\frac{\partial D}{\partial \omega_k} \to \infty$.

Formellement, l'atteinte d'une cible articulatoire n'implique pas automatiquement l'atteinte d'une cible acoustique. En effet, le système (7) montre que l'atteinte d'une cible articulatoire, $\frac{d\theta_i}{dt} = 0,\ i = 1...T$, entraîne que l'expression à droite du signe d'addition est nulle. Cela signifie que soit $\frac{d\omega_k}{dt} = 0$ (c.-à-d. que la cible acoustique est effectivement atteinte), soit $\frac{\partial D}{\partial \omega_k} = 0$. Si la dernière condition était remplie, plusieurs fréquences différentes correspondraient à des formes très semblables. Il est implausible que cette condition décrit une situation physiologiquement ou phonétiquement pertinente.

5. GESTES ARTICULATOIRES

Le geste articulatoire est la réalisation d'une constriction évolutive du conduit vocal. Les modèles gestuels précisent que le patron articulatoire élémentaire est soit un mouvement de va-et-vient d'un articulateur individuel par rapport à sa posture de repos [Lad96], soit la réalisation momentanée d'une constriction du conduit, suivi du retour à sa forme neutre [Bro86]. Afin de simplifier la notation, nous supposons ci-après qu'un geste élémentaire contrôle directement la section droite S_j d'une zone du conduit. Mais, le raisonnement qui suit s'applique également à d'autres paramètres articulatoires.

L'équation différentielle la plus simple d'un mouvement fermant ou ouvrant est l'équation logistique (8) [Tom92].

$$\frac{dS_j}{dt} = a_j(S_j - S_{1j})(S_j - S_{2j}). \quad (8)$$

Le mouvement se fait de la cible S_1 vers la cible S_2 ou vice versa en fonction du signe du paramètre a_j qui détermine la vitesse de transition. En insérant le geste (8) dans l'expression (4), nous obtenons le modèle (9).

$$\sum_{j=1}^{N} \frac{\partial D}{\partial S_j} a_j (S_j - S_{1j})(S_j - S_{2j}) + \frac{\partial D}{\partial \omega_k}\frac{d\omega_k}{dt} = 0. \quad (9)$$

Le modèle gestuel (9) est commandé par les activations et désactivations successives des cibles élémentaires S_1 et S_2. Les activations - désactivations des cibles sont fixées dans une partition [Bro86]. Le modèle (9) peut être réécrit sous une forme alternative (10), qui suggère que le mouvement des formants est dicté par une somme pondérée de gestes non-compétitifs ayant trait à des zones articulatoires différentes. La pondération est le rapport entre les deux dérivées partielles $\frac{\partial D}{\partial S_j}$ et $\frac{\partial D}{\partial \omega_k}$, pondération à laquelle la théorie quantale de la production de la parole alloue une importance particulière (cfr notre discussion ci-dessus).

$$\frac{d\omega_k}{dt} = -\frac{\sum_{j=1}^{N} \frac{\partial D}{\partial S_j} a_j (S_j - S_{1j})(S_j - S_{2j})}{\frac{\partial D}{\partial \omega_k}} \quad (10)$$

Dans le cadre de théories qui prévoient la compétition entre gestes, le modèle (8) doit être réécrit afin d'admettre des cibles multiples.

$$\frac{dS_j}{dt} = a_{1j}(S_j - S_{1j})(S_j - S_{2j}) + a_{2j}(S_j - S_{1j})(S_j - S_{3j}). \quad (11)$$

Les mouvements suivis et les cibles atteintes dépendent alors du timing des activations de plusieurs cibles virtuelles et vitesses de transitions a_{ij}, activations qui sont gérées au niveau de la partition.

6. DISCUSSION

Les développements des sections 4. et 5. suggèrent que le principe quantal d'une part et les théories fondées sur des gestes ou cibles d'autre part ne sont pas exclusifs. En effet, le principe quantal se réfère à la sensibilité aux changements articulatoires des fréquences propres. Ces sensibilités sont exprimées analytiquement dans les expressions (7) et (10) à travers un rapport de dérivées partielles qui pondèrent gestes et cibles. Par conséquent, principe quantal et gestes ou cibles peuvent entretenir des relations synergiques, car le choix des zones du conduit désignées par le principe quantal permettrait l'atteinte des cibles acoustiques avec une faible précision des cibles ou gestes articulatoires.

BIBLIOGRAPHIE

[Bon83] Bonder L. (1983), " The n-tube formula and some of its consequences ", Acoustica Vol. 52, pp. 216-226.

[Bro86] Browman C., L. Goldstein (1986), " Towards an articulatory phonology ", Phonology Yearbook, Vol. 3, pp. 219-252.

[Cio97a] Ciocea S. (1997), " Détermination de la fonction d'aire du conduit vocal par inversion acoustico-géométrique semi-analytique ", Thèse de doctorat, Université Libre de Bruxelles.

[Cio97b] Ciocea S., J. Schoentgen, L. Crevier-Buchman (1997), " Analysis of dysarthric speech by means of formant-to-area mapping ", 5th European Conference on Speech Communication and Technology, Rhodes, GrËce, pp. 2547-2550.

[Cio96] Ciocea S, J. Schoentgen, F. Bucella (1996), " A comparative study in the framework of formant-to-area mapping of continuous and discontinuous area function models of the vocal tract ", ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, Mierlo, Pays-Bas, pp. 83-87.

[Hei65] Heinz J., Stevens K. (1965), " On the relation between lateral cineradiographs area functions, and acoustic spectra of speech ", 5th International Congress of Acoustics, Liège, Belgium, Paper A44.

[Jos95] Jospa P., Soquet A., Saerens M. (1995), " Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network ", Levels in speech communication, Elsevier, Amsterdam, pp. 103-113.

[Lin63] Lindblom B.E. (1963), " Spectrographic study of vowel reduction ", The Journal of the Acoustical Society of America, Vol. 35, pp. 1773-1781.

[Lad96] Ladefoged, P., Maddieson I (1996), " The sounds of the World's Languages ", TJ Press Ltd, Padstow, Cornwall, Great Britain

[Pit96] Pitermann M., S. Ciocea, J. Schoentgen (1996), " Influence de la vitesse d'Èlocution et de l'accent sur des cibles vocaliques estimÈes aux niveaux acoustique et quasi articulatoire ", Comtes-rendus des XX^Emes JournÈes d'Etudes sur la Parole, SociÈtÈ FranÁaise d'Acoustique & European Speech Communication Association, Avignon, France, pp. 95-98.

[Sch97a] Schoentgen J., S. Ciocea (1997), " Kinematic formant-to-area mapping ", Speech Communication, Vol. 21, pp. 227-244.

[Sch97b] Schoentgen J., S. Ciocea (1997), " Post-synchronization via formant-to-area mapping of synchronously recorded speech signals and area functions ", 5th European Conference on Speech Communication and Technology, Rhodes, GrËce, pp. 1799-1802.

[Ste89] Stevens K. (1989), " On the quantal nature of speech ", Journal of Phonetics, Vol. 27, pp. 3-45.

[Tom92] Tomassone R., S. Audrain, E. Lesquoy, C. Miller (1992), " La régression - nouveaux regards sur une ancienne méthode statistique ", Inra et Masson, Paris, pp. 108-109.

A propos des degrés de liberté dans la coarticulation d'un locuteur français filmé 50 fois par seconde

VILAIN Anne, ABRY Christian, BADIN Pierre

Institut de la Communication Parlée, CNRS UPRESA 5009
Université Stendhal, DU, 1180 avenue centrale
BP 25, 38040 Grenoble cedex 9, France
Tél. : 04 76 82 41 19 - Fax : 04 76 82 43 35
e-mail : vilain,abry,badin@icp.inpg.fr - http ://www.icp.inpg.fr

ABSTRACT

Questions on speech production are here addressed on the basis of a 30-second X-ray film built on a corpus of VCV sequences uttered by a skilled French speaker. This high-quality observation, in addition to other previously recorded speakers, allows us to establish a coarticulation typology in relation to the degrees of freedom of the vocal tract. These data stand as very meaningful cues for the study of the economy and evolution of phonetic systems, starting from the emergence of coarticulation in babbling.

1. INTRODUCTION

Nous présentons ici un corpus de vues de profil du conduit vocal issues d'un film cinéradiographique, qui représente une base de données de grande valeur pour l'étude de la production de la parole, ne serait-ce que par sa rareté et sa qualité. Ce film a donné lieu à une modélisation articulatoire, le modèle Gentiane, à partir duquel il est désormais possible d'acquérir une quantité illimitée de données sans plus exposer le sujet aux rayons X.

Ces données et simulations nous permettront d'établir les contraintes de production des segments en séquences et ceci de façon à pouvoir les relier à l'organisation et à l'évolution des systèmes sonores.

Dans ce sens, nous avons appliqué ici-même la notion — venue de la robotique — de degrés de liberté en excès à un *plant* sous *contrôle*, afin d'en retirer plusieurs bénéfices : dégager les degrés de liberté de notre locuteur pour élaborer le *plant* Gentiane (section 2); observer l'exploitation de différents degrés de liberté dans la coarticulation de types articulatori-acoustiques majeurs (section 3); mettre en évidence l'émergence de nouveaux degrés de liberté au cours des processus évolutifs de la coarticulation (section 4); tester enfin une théorie de l'ontogénèse de la parole, en vue de modéliser l'émergence pas à pas des degrés de liberté dans la coarticulation du babillage canonique (section 5).

Le modèle Gentiane s'inscrit dans une modélisation "orientée sujet". Cette approche *non normalisatrice* nous permet de modéliser la génération des premiers pas de la parole, soit la morphogenèse du babillage, sur plusieurs "robots parlants". Il devient alors possible de mettre en évidence, à la fois : (i) une invariance de commande centrale cyclique implémentée périphériquement sur l'articulateur porteur mandibulaire; (ii) une variabilité radicale dans les produits résultants, s'expliquant par la conformation du sujet et les degrés de liberté de ses articulateurs, en fonction de sa stratégie individuelle. Se trouvent ainsi modélisés les premiers pas de la coarticulation, pour ainsi dire *ab ovo* .

2. MÉTHODE, DONNÉES ET MODÈLE : GENTIANE

2.1 Dispositif expérimental

Ce corpus cinéradiographique a été réalisé pour nous à l'hôpital Schiltigheim de Strasbourg, grâce à la collaboration de l'Institut de Phonétique de Strasbourg, avec la participation de Gilbert Brock, Agnès Hennel, Tahar Lallouache, Péla Simon et Jean-Pierre Zerling.

Procédure Le corpus a été enregistré simultanément en cinéradiographie par une caméra 35 mm avec un émetteur de rayons X et un amplificateur de radiance, en vidéo par une caméra et en audio par un DAT, avec des repères permettant une resynchronisation ultérieure de ces 3 enregistrements. Quatre pastilles de plomb ont été utilisées comme points de repère supplémentaires pour les images : l'une était collée sur le menton du sujet et les trois autres sur le dos, la lame et la pointe de la langue. Ces plombs apportent des informations précises et précieuses, en particulier sur les coordonnées de l'apex et la cavité sub-linguale.

Corpus Le corpus enregistré est constitué de combinaisons de séquences Voyelle-Consonne-Voyelle avec pour Consonnes [b, d, g, ʒ, v] et pour Voyelles [i, y, u, a]. Il a été établi de façon à former la "partition phonétique" qui permette, en un minimum de temps, d'explorer au maximum l'ensemble des degrés de liberté du conduit vocal.

Locuteur Le corpus de base a été produit par un locuteur unique qui, étant à la fois phonéticien et chanteur, possède une intelligence des tâches et une maîtrise exceptionnelle de son conduit vocal, pour bien mettre en évidence l'espace de travail de la parole, ceci en un temps record (30 s) afin de minimiser son exposition aux rayons X.

2.2. Traitement des données

Expertise des tracés Les images cinéradiographiques ont été projetées sur papier en chambre noire à l'aide d'un agrandisseur photo; puis les contours du conduit vocal ont été tracés au crayon à la main en taille réelle, pour obtenir

une banque de 1282 vues médio-sagittales. Il n'est actuellement pas envisageable de détecter automatiquement les contours d'une image radiographique, en raison de la superposition des ombres des différents organes. Le repérage des contours nécessite de prendre continuellement des décisions en fonction de la bonne connaissance des événements du conduit vocal. Seul un expert humain peut actuellement assumer ce rôle.

Numérisation Après scannérisation, chacun des sous-contours qui constituent le conduit vocal a été numérisé séparément de façon à permettre des mesures automatiques sur chaque articulateur. A chaque image du film cinéradiographique a ensuite été appliquée une grille de mesure bipolaire composée de 27 lignes. Les intersections des contours du conduit vocal avec chaque ligne de la grille correspondent à une série de 27 mesures qui caractérisent chaque configuration articulatoire, et fournissent les données destinées à l'élaboration d'un modèle articulatoire statistique.

2.3 Modélisation anthropomorphique individuelle

Selon nous, l'étude du polymorphisme des productions phonétiques sera rendue possible par la diversification de modèles articulatoires "orientés sujet" — soient des modèles anthropomorphiques individualisés — établis selon des méthodologies comparables. C'est pourquoi nous avons développé le modèle Gentiane [Vil97], à partir des données cinéradiographiques dont nous disposions. C'est un modèle à 8 degrés de liberté (1 pour la mandibule, 2 aux lèvres, 4 sur la langue, 1 pour la hauteur du larynx), extraits à partir d'une analyse statistique inspirée de Maeda [Mae89]. Outre le modèle de ce dernier, cette analyse a déjà éprouvée par nous pour le modèle Bergame [Bad95], élaboré à partir d'un autre locuteur. Grâce à Gentiane, nos données pourront être complétées à volonté, sans autre enregistrement cinéradiographique, ceci par une méthode de reconstruction du contour de la langue à partir de quelques points acquis par articulographie électro-magnétique sur le même sujet [Bad97].

3 DEGRÉS DE LIBERTÉ DANS LA COARTICULATION DES CONSONNES

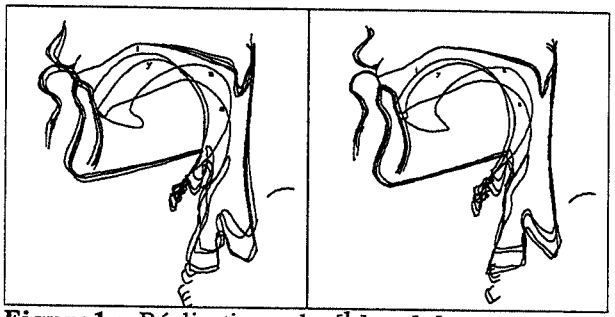


Figure1 : Réalisations de [b] et [v] en contextes [i,y,u,a]

Les réalisations de [b,d,g,ʒ,v] en divers contextes vocaliques font apparaître un certain nombre de contraintes et de degrés de liberté caractéristiques de chaque consonne. Le lieu et le mode de constriction impliquent certaines contraintes, tandis que les articulateurs non recrutés restent mobilisés par la coarticulation avec le contexte vocalique. De fait les articulations vocaliques mobilisent essentiellement la langue. Les articulations consonantiques nécessitant une constriction autre que linguale permettent donc une plus grande préservation de la position vocalique de la langue : c'est le cas du [b] et [v]. Cependant ces deux articulations labiales amènent deux types de comportement distincts.

3.1 Occlusion labiale : une constriction simple et "perméable"

Le [b] présente une constriction dont la seule contrainte semble être qu'elle soit bilabiale, les autres articulateurs restant parfaitement libres. L'espace de la rencontre des deux lèvres possède donc une certaine variabilité. Les lèvres sont même suffisamment libres pour accompagner la protrusion de [y, u], ce qui modifie le lieu de l'occlusion (figure 1).

3.2 Constriction labio-dentale : une précision "pointilleuse" et "gratuite" de la lèvre inférieure

Le [v] quant à lui présente chez notre locuteur un comportement extrêmement stable, puisque la position de la lèvre inférieure, soit la constriction labio-dentale ne varie absolument pas quel que soit le contexte (figure 1). Cette précision n'est pourtant pas imposée par des facteurs articulatori-acoustiques, puisqu'une variation du point de constriction n'apporterait aucune modification notable de la qualité acoustique de la fricative, comme c'est le cas de la légère variation de protrusion observée sur la lèvre supérieure. Il semble qu'il s'agisse ici d'une stratégie individuelle, puisque le locuteur PB, à partir duquel fut établi le modèle Bergame [Bad95], ne présentait pas ce type de précision. Ce dernier locuteur avait par contre la particularité de relever la lèvre supérieure au moment où la lèvre inférieure rejoignait l'incisive supérieure, sans doute pour l'acoustique de la friction.

3.3 Occlusion lamino-apicale : un "point" de contrainte acoustique sur la langue

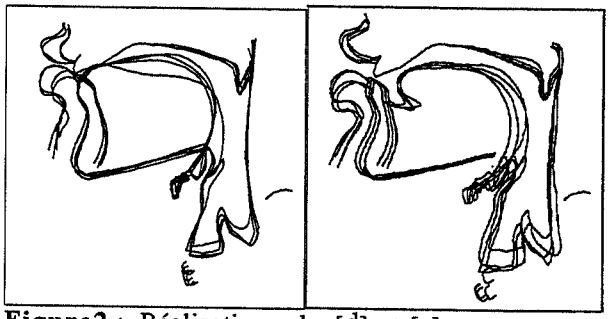


Figure2 : Réalisations de [d] et [ʒ] en contextes [i,y,u,a]

Le [d] nécessite un contact de la pointe de la langue contre les alvéoles pour obtenir une occlusion franche, c'est-à-dire qui utilise la précision de l'apex à l'occlusion et à la détente [Lab95]. Du fait du découplage de l'apex par

rapport à la masse de la langue (degré de liberté apical ou laminal), le corps de celle-ci garde une certaine flexibilité : la configuration propre à la voyelle est ainsi modifiée mais reste cependant visible (figure 2).

3.4 Constriction lamino-dorsale : une zone de contrainte acoustique encore plus forte

Le [ʒ] montre le même type de comportement, mais avec des contraintes plus fortes, puisque la constriction nécessaire à la turbulence voulue demande une grande précision, un chenal assez long sur la lame de la langue, et une tenue de cette position pendant la friction. De plus le [ʒ] optimise son acoustique par une protrusion des lèvres qui atténue, certes, la distinction entre voyelles arrondies et non-arrondies (plus visible avec le [d], figure 2), sans pourtant la mettre en danger perceptif [Abr80].

3.5 Occlusion vélaire : une coarticulation massive du corps de la langue

Le [g] est une consonne très plastique: elle présente une articulation dorsale, mobilisant la masse de la langue, qui est déjà recrutée pour la voyelle. La masse de la langue est peu mobile et possède seulement deux degrés de liberté (corps et dos), ce qui a pour conséquence une forte influence de la position vocalique sur le lieu de constriction consonantique. Les lieux d'occlusion du [g] des contextes [u] à [i] sont ainsi distants pour ce locuteur de près de 2 cm, allant de la région post-palatale à la zone vélaire (figure 3).

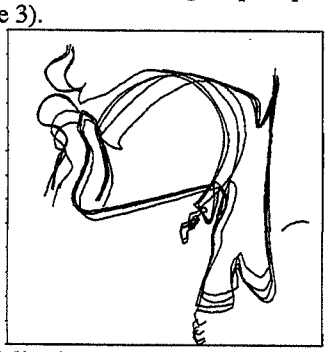


Figure3 : Réalisations de [g] en contextes [i,y,u,a]

4 DEGRÉS DE LIBERTÉ ET VIABILITÉ ARTICULATORI-ACOUSTIQUE

Une séquence dont la réalisation articulatoire est rendue difficile par le manque de degrés de liberté des articulateurs risque de se transformer selon les variations de la coarticulation. Ainsi, dans une séquence [gi], la consonne peut être modifiée par le contexte de voyelle antérieure haute au point de se palataliser (figure 4 : par comparaison avec [gu]). Puis, en raison du faible découplage des articulateurs, avoir une détente sous forme d'occlusive [ɟi] ou constrictive [ji] palatales. Mais alors une telle séquence — soit une semi-voyelle coarticulée avec une voyelle de même lieu — présente un piètre contraste acoustico-perceptif. De ce fait, ce genre de séquence est peu viable, et effectivement, elle est majoritairement évitée dans les systèmes des langues du monde [Mad93] : "[ji] and [wu] are quite consistently avoided entirely or sharply lower in occurrence than [expected]".

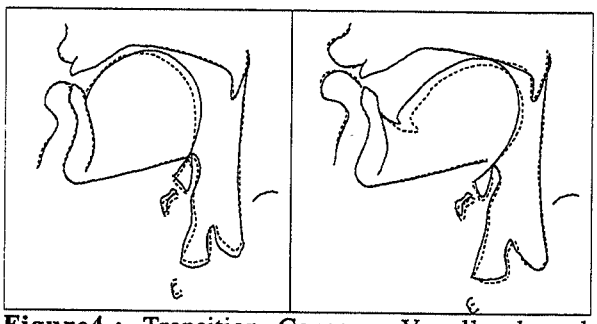


Figure4 : Transition Consonne-Voyelle dans les séquences [gi] et [gu] (La consonne est en trait plein, la voyelle en pointillés)

A ce moment de l'évolution phonétique, la viabilité contrastive de la séquence peut passer par un changement de *régime* articulatori-acoustique. En effet si l'on modifie légèrement le degré de la constriction, un phénomène de turbulence peut apparaître, et ainsi une séquence du type [ʤi] puis [ʒi]. L'exemple du français est sur ce point instructif. On peut penser que les mots commençant par [ʒi] dans cette langue proviennent du stade [ɟi], affriqué en [ʤi], avant de donner [ʒi] (comme *girolle* de *girare*). Mais l'évolution de [j] latin, devant [i] (indirectement pour *gîte* de *jacere*) et devant les autres voyelles antérieures (cf. *jam* dans *déjà* ou *jamais*), et surtout devant les postérieures (cf. *joug* de *jugum*), cette évolution montre bien qu'il y a eu optimisation du contraste par passage à la friction. Celle-ci est obtenue en exploitant la non-linéarité des relations entre les niveaux articulatoire et acoustique: on passe en régime fricatif avec de très petites variations d'aire à la constriction.

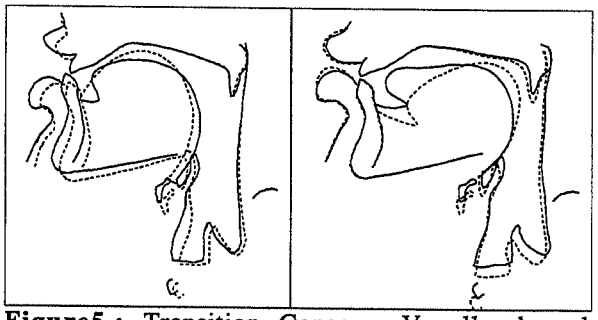


Figure5 : Transition Consonne-Voyelle dans les séquences [ʒi] et [ʒu]

La production de cette friction ne sera reprise sous contrôle véritablement actif, qu'en recrutant les degrés de liberté de l'apex ou de la lame de la langue (visible pour [ʒi] et surtout pour [ʒu], figure 5). La séquence obtenue présente un découplage des articulateurs lame/masse de la langue, bien meilleur, même pour [dʒu] par comparaison avec [ju], lequel doit avancer au maximum la position de [u] vers [j], une assimilation de lieu qui se produit en quelque sorte "dans le dos du son", en profitant : (i) de l'anticipation dans le [j] de la fermeture des lèvres (à laquelle cette constrictive n'est pas sensible, cf. [juju] arrondi de part en part), (ii) et de l'inaudibilité de l'avancée du lieu de [u] une fois les lèvres closes [Bai91]. Tout en

respectant les viabilités du système phonologique de la langue, un tel découplage assure la stabilité articulatori-acoustique du son produit et améliore notablement le contraste acoustico-perceptif de la séquence.

Comme nous l'avons déjà constaté, la qualité acoustique de ce [ʤ] ou [ʒ] sera par la suite optimisée par une articulation labialisée, qui a pour effet d'augmenter la cavité de résonance de la friction.

5 UN SEUL DEGRÉ DE LIBERTÉ OU LES PREMIERS PAS DE LA COARTICULATION

Que se passe-t-il lorsque ces différents degrés de liberté — que nous avons vu se maintenir ou disparaître à différents degrés dans la coarticulation (section 3), voire émerger pour des raisons de stabilité articulatori-acoustique (section 4) — se réduisent au seul contrôle de l'organe porteur de la langue et des lèvres : la mandibule ? Autrement dit lorsque le conduit vocal — respectivement le modèle articulatoire — ne fonctionne qu'avec un seul degré de liberté, tous les autres articulateurs habituellement actifs étant "gelés" ?

Dans cette perspective de recherche, nous avons pu commencer à tester certaines hypothèses concernant l'émergence du langage dans l'ontogenèse selon la théorie "Frame, then Content" de MacNeilage [Mac98, Abr98], basée sur l'observation du babillage canonique. Cette expérimentation a mis en évidence, à partir d'une commande centrale identique, un phénomène de variation radicale des produits périphériques, aussi différents que [bəbə] *vs.* [dædæ], selon les modèles individuels utilisés [Abr97]. Ce phénomène nous semble être un apport majeur dans l'émergence de la coarticulation, puisque la forme finale peut être considérée simplement comme le "sous-produit" de trois types de contraintes : (i) la différence de conformation des sujets ; (ii) leurs stratégies de synergie individuelles ; (iii) et finalement le "gel" de tous les autres degrés de liberté, excepté la mandibule, ce qui simule l'absence d'indépendance des articulateurs portés, à ce stade du développement (7 mois).

6 CONCLUSION : DE DEGRÉ EN DEGRÉS

C'est par la suite qu'émergeront les autres degrés de liberté, dégagés pour Gentiane (section 2), permettant à la fois le contrôle de produits acoustiquement différenciés (après 7 mois, ce qui nécessitera de développer d'au moins autant la modélisation en section 5), une coarticulation variée, dans l'ensemble relativement souple, donc fluide (grâce aux degrés de liberté observés en section 3), et une stabilité articulatori-acoustique de haute volée (par une optimisation évolutive rencontrée en section 4), soit au bout du compte une dextérité permettant de produire sur une porteuse à 6 cycles par seconde (la mandibule fonctionnant comme "pace-maker" de la syllabe), une cadence allant jusqu'à 15 modifications quantales du conduit vocal, dans le même temps: celui de la parole.

RÉFÉRENCES

[Abr80] Abry C., Boë L.-J. (1997), "À la recherche des corrélats géométriques discriminants pour l'opposition d'arrondissement vocalique en français", Labialité et Phonétique... (Abry Ed), Ellug, Grenoble, pp. 217-237.

[Abr97] Abry C., Badin P., Vilain A., Stefanuto M., Boë L.-J. (1997), "Une théorie de l'ontogenèse à l'épreuve de trois modèles articulatoires anthropomorphiques", Programme GIS Cognisciences, Rapport mi-parcours.

[Abr98] Abry C., Boë L.-J., Laboissière R., Schwartz J.-L. (1998), "Emergence of a new puzzle for the evolution of speech and language?", commentaire à [Mac98], à paraître.

[Bad95] Badin P., Gabioud B., Beautemps D., Lallouache T.M., Bailly G, Maeda S., Zerling J.P., Brock G. (1995), "Cineradiography of VCV sequences : articulatory-acoustic data for a speech production model", 15th ICA, Trondheim, Norway, pp. 349-352.

[Bad97] Badin P., Baricchi E., Vilain A. (1997), "Determining tongue articulation : from discrete fleshpoints to continuous shadow", Eurospeech, Rhodes, Greece, 1, pp. 47-50.

[Bai91] Bailly G., Laboissière R., Schwartz J.L. (1991), "Formant trajectories as audible gestures : an alternative for speech synthesis. Journal of Phonetics", 19, pp. 9-23.

[Lab95] Laboissière R., Galvan A. (1995), "Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences", ICPhS 95, Stockholm, Suède, vol. 1, pp. 358-361.

[Mac98] MacNeilage P.F. (1998), "The Frame/Content theory of evolution of speech production", Behavioral and Brain Sciences, à paraître.

[Mad93] Maddieson I. (1993), "The structure of segment sequences", UCLA Working Papers in Phonetics, 83, pp. 1-8.

[Mae89] Maeda S. (1989), "Articulation compensatoire des voyelles : analyse de données cinéradiographiques avec un modèle linéaire", Mélanges de phonétique générale et expérimentale à Péla Simon, Institut de Phonétique de Strasbourg, pp. 545-562.

[Vil97] Vilain A. (1997), Un nouveau modèle articulatoire pour la synthèse et le contrôle robotique de la parole : Gentiane. DEA Sciences du Langage, ICP, Grenoble.

Analyse et codage

CODAGE DE CANAL A SOURCE DEPENDANTE : APPLICATION AU CODEUR " *CELP* "

B. BOUDRAA*, M. BOUZID, M. BOUDRAA*, B. GUERIN**

* Institut d'électronique, USTHB, BP32 EL-ALIA, Bab Ezzouar, Alger, Algérie
** ICP/INPG, 46 AVENUE FELIX VIALLET , GRENOBLE, FRANCE

Abstract. - Distortion of parameters of a vector quantization system (QV), due to the channel noise, can be alleviated significantly by a judicious assignement of binary code-words to code-vector indexes of this system. To carry out this index assignement technique, we have developped an iteratif algorithm based on the principle of the simulated annealing. The goal of the algorithm is to find globally optimal channel-codes, destined for the implict protection of the codebook indices of a 4.8 KBPS CELP speech coder. Such source-dependant channel codes can be obtained by minimizing an appropriate distortion criterion.

Mots clés. Codage source-canal, Assignation d'indice, Recuit simulé, Codeur *CELP*

1. Introduction.

Dans cet article, on expose une méthode récente de codage de canal à source dépendante (*CCSD*). On utilise une technique d'assignation d'indice afin de mettre en oeuvre ce codage de canal. Cette technique est utilisée sans ajout de redondance. De tels codes de canal qui dépendent du codeur-source sont obtenus en minimisant un critère d'erreur approprié à notre problème d'assignation d'indice.

Afin de mettre en oeuvre cette technique d'assignation d'indice, nous avons développé un algorithme itératif basé sur le principe du *RS* (Simulated Annealing). Pour valider l'efficacité de ce *CCSD*, dans le cas des transmissions à travers un canal bruité, nous avons implanté un codeur de parole de type *CELP* (Code Exited Linear Prediction) [MOR 91]. Ce codeur à 4.8 KBPS opère sous des conditions spécifiées d'erreur de transmission.

2. Modèle d'une QV dans le cas d'un canal bruité.

Un quantificateur vectoriel Q est une fonction d'un espace euclidien R^k de dimension k vers un sous-ensemble fini Y (dictionnaire de taille $L = 2^n$, où n représente le nombre de bits par indice binaire) de R^k, qui affecte à un vecteur-source x un vecteur-code correspondant y de même dimension (k). Il est donné par :

$$Q : R^k \rightarrow Y$$
$$x \rightarrow y_i \qquad (1)$$

Avec $Y = \{y_0, y_1, ..., y_{L-1}\}$ et $y_i \in R^k$.

Dans le cas des systèmes à *QV* conçus pour des transmissions à travers un canal non bruité, le codage canal-source sans redondance par la technique d'assignation d'indice peut fournir une robustesse contre les erreurs de canal, dans le cas où ces systèmes sont appelés à opérer dans un environnement perturbé [POT 95], [ZEG 90], [CHI 95], [KLE 92], [FAR 90]. La fonction d'assignation d'indice affecte à chaque vecteur-code d'indice $i \in \{0,1\}^n$ ($\{0,1\}^n$ ensemble des L indices binaires perturbé [POT 95], [ZEG 90]), un mot-code binaire unique, appartenant au même ensemble. On symbolise cette configuration par b ($b \equiv (b(0),...,b(L-1))$, où $b(i)$ est le mot-code attribué à l'indice binaire i, c.-à-d., au vecteur-code y_i.

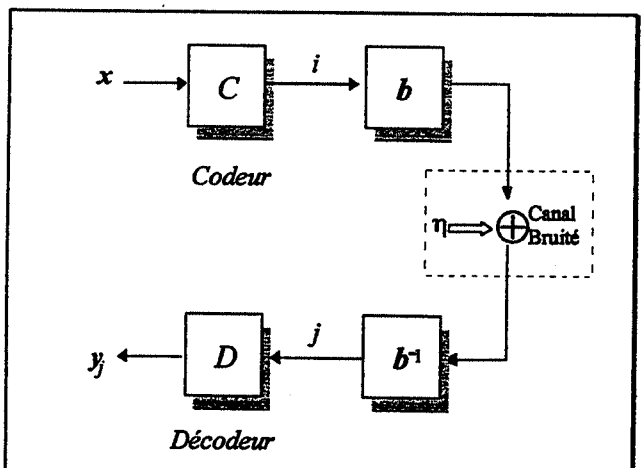


Fig. 1 : Schéma-bloc d'un système à *QV* pour des transmissions par canal bruité

η est une variable aléatoire dont les valeurs appartiennent à l'ensemble $\{0,1\}^n$ et b^{-1} est la fonction inverse de b.

3. Définition du critère d'erreur

La distorsion causée par la représentation d'un vecteur-source x de R^k par un vecteur-code y est définie par une certaine mesure de distorsion $d(x, y)$. La distorsion moyenne totale causée par l'effet combiné du

quantificateur vectoriel et du bruit de canal est donnée par [CHI95], [FAR 90] :

$$\varepsilon_T = \frac{1}{k}\sum_{i=1}^{L}\sum_{j=1}^{L} p(b(j)/b(i)) \int_{R_i} p(x).d(x,y_j).dx \qquad (2)$$

où $p(x)$ est la fonction densité de probabilité du vecteur-source et $p(b(j)/b(i))$ est la probabilité de transition du canal.

En utilisant une fonction de distorsion quadratique moyenne (MSE), le problème de la minimisation de la distorsion totale peut être réduit à la simple minimisation de la distorsion du canal ε_C, [CHI95], [FAR 90] [ZEG 95] :

$$\varepsilon_C = \frac{1}{k}\sum_{i=1}^{L} p(i) \sum_{j=1}^{L} p(b(j)/b(i)) \cdot d(y_i,y_j) \qquad (3)$$

où $p(i)$ dénote la probabilité d'occurrence d'un vecteur-code y_i.

Supposons que le canal utilisé pour la transmission des mots-codes est un canal binaire, symétrique (CBS) de probabilité de transition ε. Pour simplifier le problème du codage de canal, nous supposerons que la probabilité de transition du canal est suffisamment petite ($\varepsilon \ll 1$) afin de considérer le cas d'un seul bit d'erreur dominant par mot-code bruité [CHI 95], [POT 95] [ZEG 95]: Ainsi, le critère d'erreur approprié à l'optimisation des mots-codes est donné d'une manière simplifiée par :

$$\varepsilon_C = \sum_{i=0}^{L-1} p(i) \sum_{m=1}^{n} d(y_i, y_{b^{-1}(f(b(i),m))}) \qquad (4)$$

$f(i, m)$ est l'indice du vecteur-code obtenu par inversion (bruitage) du *m ème* bit de l'indice *i*.

4. Assignation d'indice par *RS*

La recherche de la fonction d'assignation d'indice, $b(.)$, qui minimise le critère d'erreur est un problème d'optimisation combinatoire. Il est souvent impraticable d'évaluer les performances pour toutes les configurations possibles ($L!$ assignations d'indices possibles à L vecteurs-codes). Ainsi, une technique sous-optimale doit être utilisée pour résoudre ce problème. Le *RS* est un algorithme d'optimisation puissant. Cet algorithme à caractère itératif est très utilisée pour trouver des solutions optimales à une variété de problèmes d'optimisation combinatoire [KIR 83]. La méthode du *RS* est fondée sur une analogie entre la recherche d'un état d'énergie minimale, pour un système physique, et la minimisation d'un critère d'erreur, pour un problème d'optimisation combinatoire [KIR 83]. Chaque assignation d'indice $b(.)$ simule un état d'un système physique avec un critère d'erreur ε_C qui représente l'énergie du système. L'algorithme donné ci-dessous décrit les étapes de base de notre algorithme du *RS*. La définition de la loi de décroissance de la tepérature T est un point très important. Elle permet de piloter le processus d'optimisation par RS. Plusieurs formules ont été utilisées pour définir la loi de baisse de la température. Nous retenons particulièrement la loi géométrique suivante :

$$T_k = \alpha T_{k-1} \qquad 0 < \alpha < 1 \qquad (5)$$

Etape 1: *Initialisation :*
- définir une température fictive T et l'initialiser à une valeur haute : Ti ,
- choix de l'état initial b par l'algorithme de l'amélioration itérative.

Etape 2: *procédure de Metropolis* :
- choisir aléatoirement un état b' : perturbation de l'état b et calculer $\Delta\varepsilon = \varepsilon_C(b') - \varepsilon_C(b)$

Règle d'acceptation de Metropolis:

a) si $\Delta\varepsilon < 0$, remplacer b par b', aller à l'étape 3.
b) si $\Delta\varepsilon \geq 0$, remplacer b par b' avec la probabilité $\exp(-\Delta\varepsilon / T)$, aller à l'étape 3.

Etape 3: *Test équilibre thermodynamique local*

- si dans l'étape 2 le nombre de baisse d'énergie dépasse un nombre prescrit ou si trop de perturbations, qui n'engendrent aucune baisse d'énergie, se produisent aller à l'étape 4.
- Sinon, retour à l'étape 2.

Etape 4: *Programme du recuit* :

- diminuer lentement la température T.

Etape 5: *Test de fin* :
- si T est inférieure à une certaine température de congélation prescrite : Tc ou s'il apparaît qu'un état stable est obtenu, alors fin du programme (avec b qui décrira l'état stable du système).
- Sinon, retour à l'étape 2 .

5. Description du codeur CELP

Dans l'implantation du codeur *CELP*, Nous utilisons une excitation d'ordre 2. Elle est modélisée par une combinaison linéaire de deux vecteurs-codes, sélectionnés selon une procédure d'analyse par synthèse en minimisant un critère d'erreur perceptuel [MOR 91]. Le premier vecteur, d'indice de retard $j(1)$, est extrait d'un dictionnaire prédictif de 128 vecteur et le second (d'indice statistique $j(2)$) d'un dictionnaire d'excitation statistique de même taille.
Les paramètres qui sont transmis par ce codeur au canal sont : les coefficients de prédiction LARi, les indices j(1) et j(2) ainsi que les gains gj(1) et gj(2). Le codage des

coefficients LARi est les gains de pondération est réalisé par une méthode de quantification scalaire. Les indices prédictifs et statistiques j(1) et j(2) sont codés directement en binaire par une QV. Le tableau 1 montre l'allocation des bits des paramètres à transmettre par fenêtre d'analyse de 30 ms.

Tableau 1 : Allocation des bits.

Paramètres	bits alloués	débit (bits/s)
10 coefficients LARi	6,6,5,5,5,5,4,4,4,4	1600
$4 \times j(1) + 4 \times g_{i(1)}$	$4 \times 7 + 4 \times 5$	1600
$4 \times j(2) + 4 \times g_{i(2)}$	$4 \times 7 + 4 \times 5$	1600
Total	144	4800

5. Résultats et évaluations.

Les performances de notre codeur *CELP* sont évaluées par des tests objectifs, en utilisant le rapport signal/bruit segmental (*RSBseg*). Notons que, l'application du *CCSD* est restreinte aux paramètres codés par une *QV*, à savoir les indices de retard et les indices statistiques. Des résultats expérimentaux, obtenus pour différents taux d'erreur de transmission, sont présentés.

Les performances du *CCSD* destiné à la protection des indices des dictionnaires du codeur *CELP* sont évaluées d'une manière séparée. C'est-à-dire que, nous codons et perturbons, volontairement, un type d'indice à la fois. Les résultats objectifs, donnés par la figure 2, sont obtenus à partir d'un signal de parole d'environ 33 secondes. Il est constitué par 11 séquences, prononcées par 4 locuteurs masculins et 2 locuteurs féminins. Ces résultats indiquent un bon comportement du codeur *CELP* dans un environnement bruité, où le taux d'erreur est élevé (0.1). Nous constatons aussi que le codeur présente une performance remarquable pour des taux d'erreurs faibles.

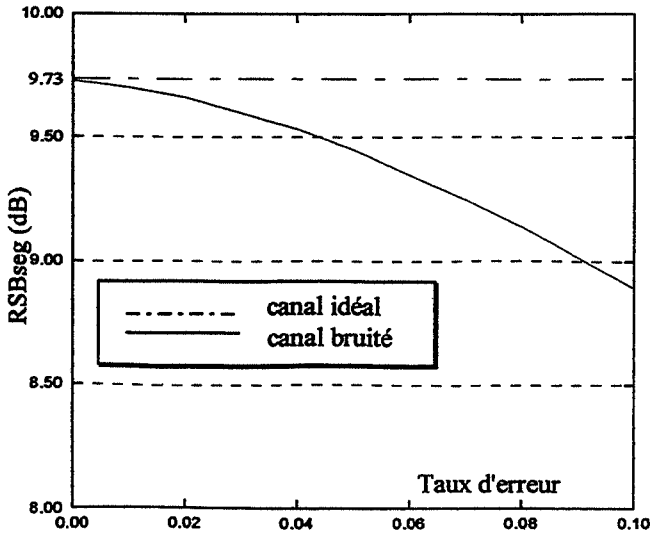


Fig.2 : Performances objectives du codage de canal des indices statistiques.

Les performances objectives du codage de canal des indices de retard sont donnés par la figure 3. Notons que ce codage de canal doit être mis à jour chaque fois que le contenu du dictionnaire prédictif change. Ceci est dû à la nature adaptative de ce dictionnaire. Ces résultats sont obtenus à partir d'un signal de parole d'environ 13 secondes. Il est constitué de 4 séquences prononcées par 2 locuteurs masculins. En examinant ces résultats, nous constatons que le codeur CELP (avec codage de canal des indices de retard) présente une bonne performance objective.

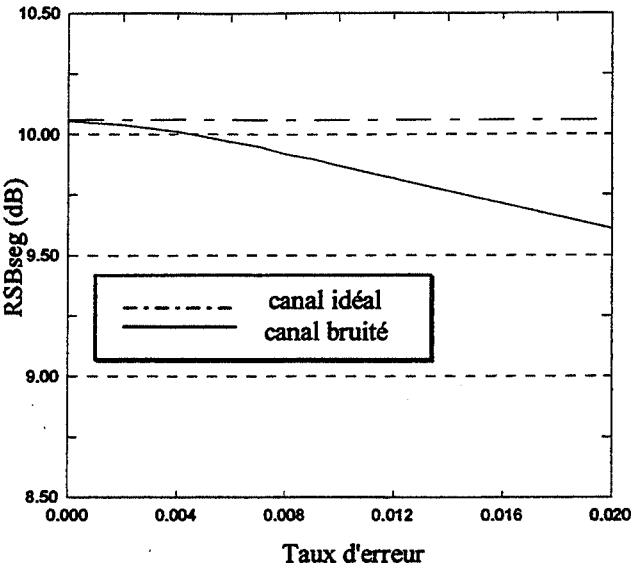


Fig.3 : Evolution du *RSBseg* dans le cas du codage de canal des indices de retard.

La figure 4 donne un exemple de résultat obtenu lors de la transmission d'un signal de parole, d'environ 3 secondes, à travers un canal bruité. Dans de cette expérience, un indice de type inconnu est altéré toutes les 8 fenêtres d'analyse. Le taux de bruitage enregistré à la fin de cette transmission est de 0.03. Le *RSBseg* moyen obtenu dans le cas de la transmission idéale est de **10.59 dB**. Pour la transmission bruitée (c), le *RSBseg* a diminué seulement de **0.05 dB**. Il est de **10.54 dB.**

7. Conclusion.

En utilisant comme exemple d'application l'algorithme du codeur CELP, nous avons montré que le CCSD, par la technique d'assignation d'indice, peut fournir une bonne protection d'erreur aux indices des dictionnaires du codeur CELP. Une amélioration significative des oerformances objectives du codeur CELP en présence de bruit de canal a été apportée par cette technique, sans avoir recours à des techniques de codage de canal redondant.

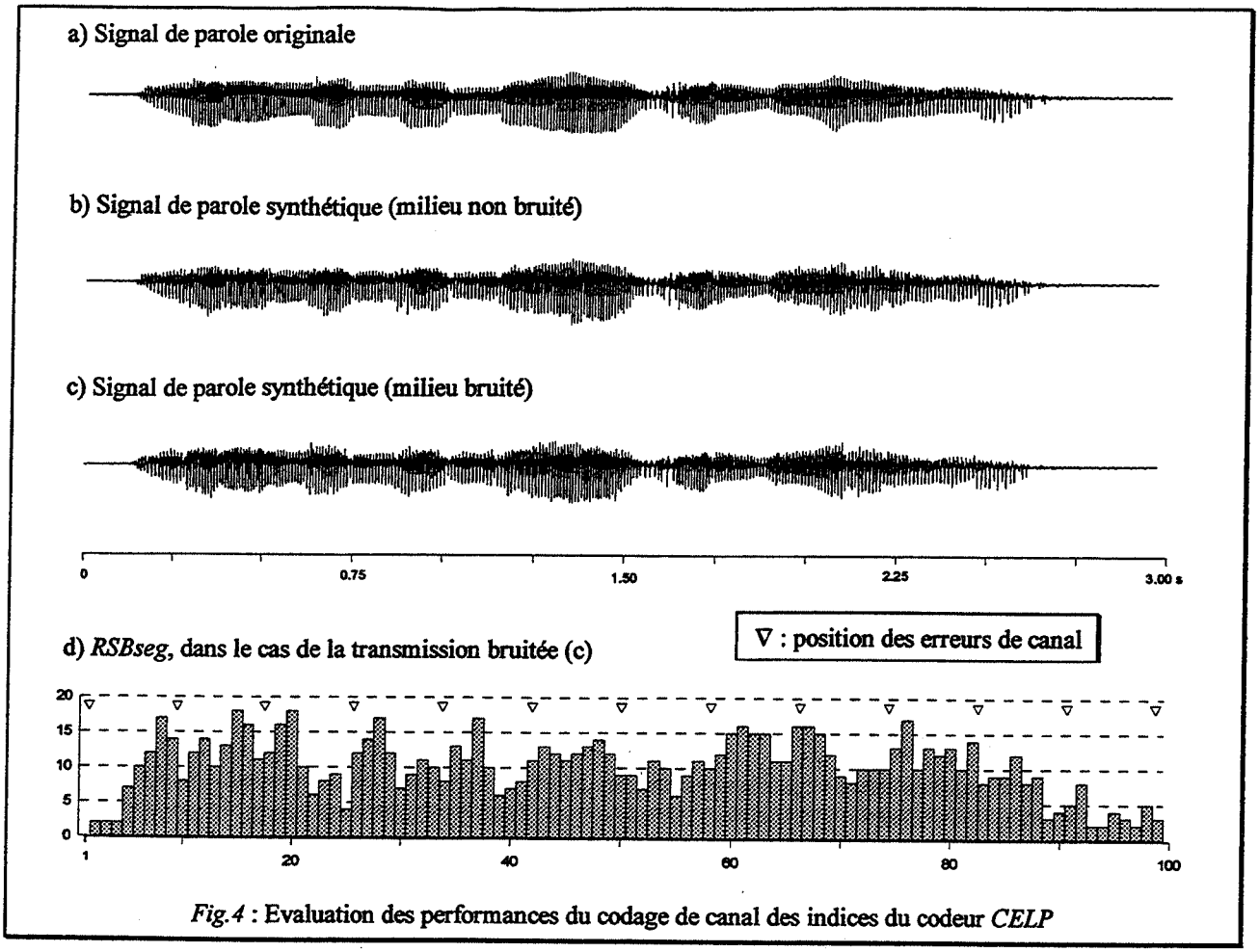


Fig.4 : Evaluation des performances du codage de canal des indices du codeur *CELP*

BIBLIOGRAPHIE

[MOR 91]

MOREAU (1991) " Codage prédictif du signal de parole à débit réduit: une présentation unifiée", Annales des télécommunications, vol 46, N° 3-4, 1991.

[POT 95]

L.C. POTTER & D.M. CHIANG (1995) , "Minimax Nonredundant Channel Coding", IEEE Trans on Communications,Vol.43, N°.2/3/4, pp. 804-811, Feb/ Mar /April 1995.

[ZEG 90]

K.ZEGER & A.GERSHO (1990), "Pseudo-Gray Coding", IEEE Trans on Communications, Vol. 38, pp. 2147-2158, December 1990.

[CHI 95]

D.M. CHIANG & L. C. POTTER "Vector Quantization For Noisy Channels : A Guide To performance And Computation ", IEEE. Trans on Circuits and systems for Video Technology, November 1995.

[KLE 92]

W. KLEIJN & R. SUKKAR (1992), " Efficient Channel Coding for CELP using Source Information ", Speech Communication, Vol. 11, N°. 6, pp 547-566, December 1992 .

[FAR 90]

N . FARVARDIN (1990), "A Study of Vector quantisation for Noisy Channels ", IEEE Trans on Information Theory, Vol. 36, N°.4, July 1990.

[KIR 83]

S. KIRKPATRICK, C. GELATT & M. VECCHI (1983) "Optimisation by simulated Annealing".Science, Vol. 220, N°. 4598, pp 671-680, 13 May 1983.

FILTRAGES LOGIQUES ADAPTES AU PITCH

M.BOUDRAA*, H.SAYOUD*, B.BOUDRAA*, B.GUERIN**
*Institut d'électronique, USTHB, BP 32 EL-ALIA, Bab Ezzouar, ALGER, ALGERIE.
**I.C.P./ I.N.P.G., 46 AVENUE FELIX VIALLET, GRENOBLE, FRANCE.
*Tel: (00.213.2) 51 55 75 *Fax (00.213.2) 51 50 39 *Telex: 64.343

Abstract

We propose some new algorithms of post-processing, called the logical filtering, destined to correct the errors of pitch detection with an optimal correction. The post-processings, thus developed, were tested in a noisy environment at - 6dB of SNR. We have, then, valued objectively the improvement in precision of each of them, by elaborating some new error rates well adapted for this survey, which we have called: EAP rate and ES3D rate. The statistical results obtained after the processing of 60 Arabic sentences, unveiled the excellent correction brought by the logical filtering (ALF). This one allowed to reduce the detection error by about 50%.

Mots Cles

Pitch, Post-Traitement, Erreur de détection.

1. Introduction

De nos jours les **PDA's (Pitch Detection Algorithms)** ont vu une diversité telle que l'on ne peut opter pour l'un d'eux avec un critère sévère justifiant son utilisation ou son choix [Bou94]. De plus, les nouveaux PDA's implantés en laboratoire sont souvent moins fiables dès que l'on traite une voix etrangère à la base de données initiale.

L'une des causes prépondérantes de cette imperfection est la large étendue spectrale du pitch qui varie de 80 Hz à 360 Hz en moyenne. Cette large étendue spectrale impose un équilibrage difficile des seuils du PDA qui est souvent non optimal. L'équilibrage aura, alors, des Maxima locaux en précision (dans certaines plages fréquentielles); cependant, dans d'autres régions fréquentielles on verra des défaillances.

Afin d'eclairer les utilisateurs de PDA's, soucieux d'une bonne précision de calcul, on a mis au point un nouvel analyseur d'erreur pour PDAs appelé ES3D ou "Erreur Spectrale en 3 Dimensions" [Bou97] et qui représente la dispersion de l'erreur sur Fo en fonction de l'erreur relative et de la fréquence du pitch. Ce volume 3D nous a montré maintes informations, tantôt sur la précision et tantôt sur les plages des fréquences où l'on constate une défaillance du PDA, appelées "***fréquences à problème***".

En utilisant ces outils statistiques, nous avons pu raffiner les courbes mélodiques (ou courbes de variation de Fo) grâce à l'élaboration d'un nouveau post-traitement appelé filtrage logique. Ce post-traitement a apporté une nette amélioration dans la base de données des résultats. Ces techniques de correction jouent le rôle de filtre contre les déformations causées par les harmoniques et les formants [Bou96], en corrigeant, ainsi, les erreurs obtenues lors de la détection.

2. Principe du Filtrage Logique

Tout signal aléatoire d'origine physique, présente une certaine continuité dans le temps. De plus cette théorie s'applique bien dans le traitement du signal, et en particulier en détection du fondamental. Dans le cas de la parole, on sait que les cordes vocales ne peuvent changer brusquement de fréquence en un laps de temps de 20 ms, ce qui confirme la théorie précédente en détection de mélodie. Ceci nous a permis d'appliquer des méthodes de filtrage et de lissage bien adaptées, à la fin de chaque détection. Toutefois, certaines méthodes ont donné des resultats non satisfaisants lors des tests pratiques en laboratoire. Ces méthodes ont, alors, été éliminées, et les méthodes sélectionnées sont celles qui améliorent nettement la précision de nos détecteurs implantés.

2.1. Filtrage Anti-Harmonique

2.1.1. Erreur Harmonique

Dans certains cas, la première composante harmonique prédomine sur le fondamental, d'où l'on obtient une fréquence dédoublée (i.e. Fo = 2xPitch). L'erreur ainsi commise s'appelle erreur harmonique.

Pour y remedier, on procède à une correction logique appropriée.

2.1.2 Algorithme

Les étapes de ce filtrage sont :

- Calcul de la fréquence moyenne :

$$(1) \qquad Fmoy = \frac{1}{N} \sum_{i=1}^{N} Fo[i]$$

avec N représentant le nombre des fenêtres analysées,

- Comparaison de Fo[i] à Fmoy, pour tout i compris entre 1 et N,
- Si 1.8 Fmoy < Fo[i] < 2.2 Fmoy, alors :

$$(2) \qquad Fo_{flt}[i] = \frac{Fo[i]}{2} \; ;$$

où $Fo_{flt}[i]$ est la valeur filtrée de Fo,

- Sinon garder l'ancienne valeur de Fo[i].

2.2. Filtrage Anti-Double-Période

2.2.1 Double-Période

Quand les amplitudes des pics du signal original sont irrégulières dans de larges proportions, alors il se peut que l'un des pics faibles soit considéré comme étant un bruit additif et sera donc dénié. Voir l'exemple illustré sur la figure 1 ci dessous. La valeur detectée sera divisée par 2 (i.e. Fo = Pitch /2).
Il s'agit, alors, d'une double période.

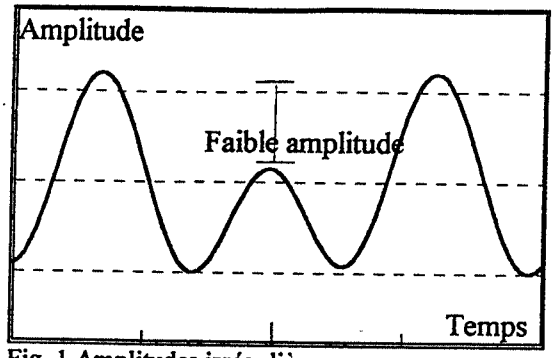


Fig. 1 Amplitudes irrégulières.

2.2.2. Algorithme

Pour y remédier, on procède à une correction particulière, définie par l'algorithme ci-dessous:

• Calcul de la fréquence moyenne Fmoy:

$$Fmoy = \frac{1}{N}\sum_{i=1}^{N} Fo[i],$$

avec N représentant le nombre des fenêtres analysées,

• Comparaison de Fo[i] à Fmoy, pour tout i compris entre 1 et N,

• Si 0.4 Fmoy < Fo[i] < 0.6 Fmoy, alors

(3) $Fo_{flt}[i] = 2xFo[i]$,

où $Fo_{flt}[i]$ étant la valeur corrigée de Fo[i].

2.3. Filtrage d' Ilôts

Rarement, des pics anormaux apparaissent sur la courbe de Fo [Sel91], Ces pics correspondent à une discontinuité brutale de passage d'un état de périodicité à un état de non périodicité (et vice-versa), puis retour à l'état initial brièvement, comme on peut le constater sur le tableau 1.

Tableau 1 Effet du filtrage d'ilot.

Fenêtre	Fo	Fo filtré
1	0	0
2	124	0
3	0	0
4	0	0

Les tests accomplis dans ce filtrage sont les suivants:

• Si Fo[i-1] = Fo[i+1] = 0 et Fo[i] ≠ 0 , alors on pose Fo[i] = 0

et • Si Fo[i-1] ≠ 0, Fo[i+1] ≠ 0 et Fo[i] = 0 , alors on pose

(4) $$Fo[i] = \frac{(Fo[i-1] + Fo[i+1])}{2}$$

Dans le cas de la parole , on admet que pendant 10 ms, il est impossible d'avoir un passage d'un voisement à un non voisement, puis retour à l'état initial (et vice-versa).

2.4. Filtrage Anti-Pic

Dans le cas de la parole, on estime que la fréquence fondamentale instantannée, pour une voix normale, non chantée, se situe dans une plage bien determinée et qui est de l'ordre de [0.64 Fmoy , 1.68 Fmoy] (résultats expérimentaux), où Fmoy représente la fréquence moyenne d'une phrase equilibrée [Kru91], courte et non chantée.
Ceci étant, on admettra qu'on a une erreur à chaque fois que Fo se situe à l'extérieur de cet intervalle. Par suite, une correction moyennée sera appliquée (techniques de lissages).

2.5. Filtrage de Continuité

Les lois physiques et acoustiques interdisent toute variation brutale de la mélodie en 20 ms. Par variation brutale, on fait allusion à une discontinuité de Fo, de 20 % ou plus, au niveau de deux trames voisines (une discontinuité de 20 % signifie une variation relative de Fo de l'ordre de 20 %, entre deux trames voisines). Un exemple de discontinuité double est illustré sur la figure 2 ci dessous.

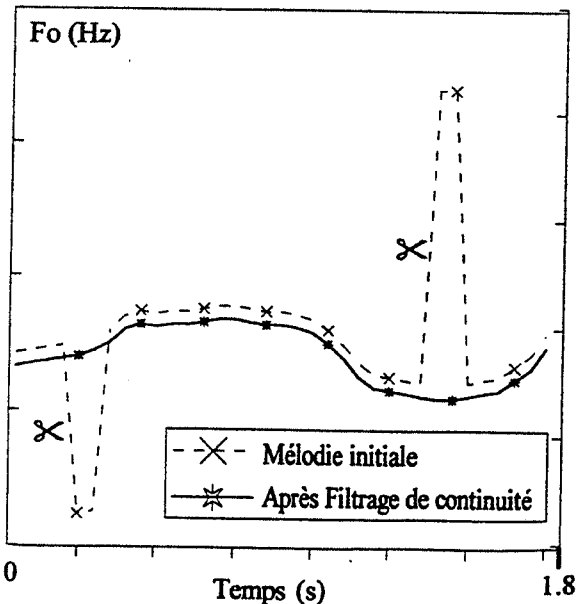


Fig. 2 Discontinuité mélodique double et sa correction par le filtrage de continuité.

Ceci peut être corrigé si, après avoir executé le suivi dynamique, on éssaie de changer, une à une, les valeurs de la discontinuité par leurs candidats secondaires du suivi. Nous procédons, alors, aux tests suivants:

• Si Candidat 1 présente une discontinuité :
alors Tester Candidat2 ;

• Si Candidat 2 présente une discontinuité :
alors Tester Candidat3.

Souvent cette discontinuité à 20 % est corrigée avec succès. Notons, quand même, que leur correction est mieux réussie pour le sexe féminin ; d'ailleurs les cas masculins présentent rarement des discontinuités (résultat expérimental).
En pratique, on fait deux filtrages de sens contraires, l'un dans le sens direct et l'autre dans le sens inverse.

3. Base de Données Parlée

Nous avons enregistré une base de données parlée, constituée de 60 phrases phonétiquement équilibrées [Bou92-JEP], prononcées par 3 locuteurs masculins et 3 locuteurs féminins. Il s'agit d'un arabe standard. La fréquence d'échantillonnage est de 10 KHz et l'analyse se fait par fenêtres à 20 ms avec un recouvrement de 50 %. Le PDA utilisé est la méthode d'AMDF [Bou94]. Afin de destabiliser le PDA et d'augmenter le taux des grosses erreurs [Kru91], on a ajouté un certain niveau de bruit blanc gaussien additif. Le RSB obtenu est, alors, de -6 dB. (le choix de -6db a été fait après avoir remarqué que les taux d'erreur commençaient à accroitre brusquement à cet RSB.)

4. Algorithme de Détection du Pitch

Le PDA utilisé est la méthode d'AMDF [Sel91], car cette méthode a, relativement, donné de bons résultats, comparativement à d'autres méthodes testées dans les mêmes conditions [Sel91], [Kru91].
Son algorithme est basé sur la formule ci-dessous:

$$(5) \qquad AMDF(k) = \frac{1}{N}\sum_{n=1}^{N-k}|s(n)-s(n+k)|$$

où k varie de 1 à M (nombre de points de l'AMDF) et N est la largeur de la fenêtre d'analyse du signal de parole s(k).
De plus, en utilisant le coefficient de confiance C_f défini par:

$$(6) \qquad C_f = \frac{Max(AMDF)}{Min(AMDF)}$$

nous pouvons rencontrer 3 cas:

A- Si $C_f > 1600$ alors on garde un candidat sûr unique:

$$(7) \quad Fo = \frac{Fe}{k_{min}}$$ (kmin = position du 1er minimum).

B- Si $C_f < 250$, alors la fenêtre est non voisée.

C- Si $250 < C_f < 1600$, alors 3 candidats probables sont mémorisés (pour le suivi dynamique).

L'étape suivante sera le post-traitement.

5. Résultats

5.1 Rendement en Précision

Afin de mesurer le rendement en précision apporté par nos post-traitements, on a calculé les taux d'erreur, avant et après le post-traitement ; ces erreurs sont calculées de la façon suivante [Bou96]:

i) Calcul de l'erreur relative entre la valeur du pitch électro-glotto-graphique [Bou92-SSA] et la valeur de Fo estimée :

$$(8) \qquad er_i = \frac{Fo_{ref_i} - Fo_i}{Fo_{ref_i}}$$

ii) Calcul de l'EAP (Erreur Absolue sur le Pitch) par :

$$(9) \qquad EAP = \frac{1}{N}\sum_i |er_i|$$

où N = Nombre de trames voisées traitées,
et Fo_{ref} = Valeur effective de Fo trouvée par la méthode Electro-Glotto-Graphique [Bou92-SSA].
Les taux d'erreur, ainsi définis, sont illustrés sur la figure 3 suivante.

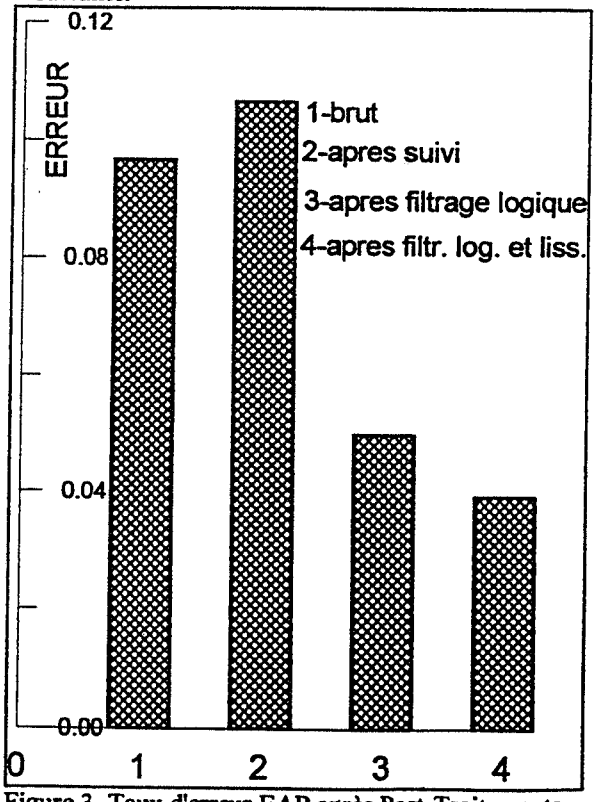


Figure 3 Taux d'erreur EAP après Post-Traitements.

Nous pouvons, alors, voir que le filtrage logique (colonne 3) a considérablement réduit le taux d'erreur EAP, relativement aux deux autres post-traitements utilisés en parallèle. A titre indicatif, le filtrage logique a permis de réduire le taux d'erreur de plus de 50%, grâce à ce dernier seulement.

5.2 Analyse de l'Erreur

Dans le but de voir l'effet de ces post-traitements dans la dispersion globale de l'erreur de détection, nous avons développé un nouvel analyseur d'erreur [Bou97], pouvant

visualiser plusieurs propriétés de ces erreurs sur un volume à 3 dimensions. Nous l'avons appelé "ES3D", c'est à dire Erreur Spectrale à 3 Dimensions. Sur la figure 4 on peut voir trois dimensions, définies par:

- **Le Taux d'erreur en % (verticalement)**
- **La bande spectrale en Hz (horizontalement)**
- **la dispersion statistique de l'erreur (diagonalement en hauteur).**

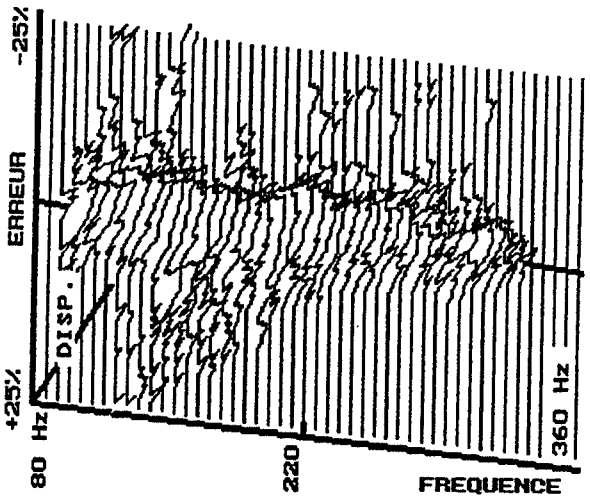


Fig. 4 Erreur Spectrale 3D du PDA AMDF, après post-traitement.

5.3. Interprétation

L'ES3D est relativement modérée donnant un aspect de PDA assez robuste avec des fréquences à problème autour de 110 Hz, ces erreurs sont dues aux harmoniques. Par contre, pour les fréquences féminines, le volume ES3D est très mince en se penchant légèrement vers le haut, indiquant une legère diminution de la fréquence fondamentale. Ceci est probablement dû aux erreurs de double période qui persistent encore apres filtrage logique. Ainsi donc, les erreurs harmoniques n'ont été que partiellement éliminées, tandis que les doubles périodes sont mieux filtrées.

Conclusion

Ayant conçu de nouveaux post-traitements adaptés au pitch (appelés Filtrages Logiques), nous avons entrepris une analyse de la précision en milieu bruité à -6 dB de RSB, en parallèle avec d'autres types de post-traitemebnts. Chacune des étapes du filtrage a contribué à l'amélioration de la précision en renforçant la stabilité du PDA.

A titre indicatif, la réduction du taux d'erreur "EAP" s'élève à environ 50% grâce à l'emploi du Filtrage logique seul, sachant qu' à -6 dB le bruit est plus puissant que le signal utile. Ceci confirme l'importance capitale de ce dernier. D'autres techniques de post-traitement telles que les techniques de lissage, se sont avérées moins importantes, mais renforcent considérablement la précision grâce à la bonne élimination des grosses erreurs (taux des trames pour lesquelles l'erreur > 10 %). Son rendement en précision n'excède pas 10%.

Finalement, on peut noter que ces techniques de filtrages sont meilleures pour le cas féminin que pour le cas masculin, ce qui se traduit par une meilleure correction des algorithmes dans le cas des erreurs de double-période.

Références

[Bou97]

M.BOUDRAA, H.SAYOUD, B.BOUDRAA, B.GUERIN (1997). Un Nouvel Analyseur d'Erreur pour PDA: "L'Erreur Spectrale 3D", Actes du 4ème Congrès Français d'Acoustique, pp 377-380, Marseille 14-18 Avr. 97.

[Bou96]

M.BOUDRAA, H.SAYOUD, B.BOUDRAA, B.GUERIN (1996). "Interprétation des erreurs de pitch de parole corrompue analysée sur des segments pseudo-stationnaires". JTEA'96, pp 369-374, Nabeul, TUNISIE.

[Bou94]

M.BOUDRAA, H.SAYOUD, B.BOUDRAA, B. GUERIN (1994). Etude comparative de plusieurs détecteurs de périodicité du signal de parole en présence de sept niveaux d'un bruit gaussien. International C.C.S, Sept. 94 ALGER , pp. IV-32 à IV-36.

[Bou92-JEP]

M.BOUDRAA, B.BOUDRAA, B.GUERIN (1992). Mise en place de phrases arabes phonétiquement équilibrées, XIX ème JEP, Bruxelles, Mai 1992.

[Bou92-SSA]

M.BOUDRAA et Al (1992). Extraction de la fréquence fondamentale du signal de parole: Approche optimale évaluée par laryngographe (EGG), Conférence SSA' 92, Blida Dec 92.

[Kru91]

D.A.KRUBSACK & R .J.NIEDERJOHN (1991). An autocorrelation Pitch Détection and Voicing Decision with Confidence Measure Developed for Noise Corrupted Speech, IEEE Trans. on acoustics, speech, and Signal Processing. Vol. ASSP-39, No 2, pp 319-541, Feb. 1991.

[Sel91]

S.A.SELOUANI (1991). Contribution à l'extraction des paramètres prosodiques du signal de parole : Cas de la fréquence fondamentale, Magister thesys, USTHB ALGIERS 1991.

Critère Statistique pour l'Automate de Détection Bruit/Parole

Lamia Karray et Jean Monné

FT.CNET/DIH/DIPS
2, Av. P. Marzin, 22307 Lannion Cedex, France
E-mail : lamia.karray@cnet.francetelecom.fr
jean.monne@cnet.francetelecom.fr

ABSTRACT

The use of speech recognition systems in noisy environments requires robustness to adverse conditions. An efficient detection of speech/non-speech segments is therefore necessary. Several approaches have been proposed in order to improve the robustness of speech/non-speech detection used for isolated word recognition in noisy conditions [1, 2, 3, 4]. In this paper, we describe a robust speech/non-speech detection algorithm based on the estimation of noise statistics: mean and variance. Results of several experiments carried out on a database collected over the GSM network show that this new approach improves the recognizer's global performances, especially in very noisy environments.

1. INTRODUCTION

Dans le contexte de reconnaissance de la parole à travers le téléphone, les performances de reconnaissance dans des Serveurs Vocaux Interactifs (SVIs) se trouvent dégradées par rapport à la reconnaissance de parole directe (non téléphonique). Cette limitation est due au fait qu'en plus des variabilités intra- et inter-locuteurs (e.g., accent, stress, effet Lombard, ...), le signal de parole est modifié par le bruit ambiant et les distorsions fréquentielles dues à l'éventuel codage et aux perturbations des canaux de transmission qui véhiculent la parole. Le bruit ambiant devient d'autant plus présent et gênant que, de nos jours, un nombre croissant de communications se font de l'extérieur : à partir d'une cabine téléphonique ou avec un téléphone mobile.

Quand les mots du vocabulaire sont noyés dans le bruit, leur détection devient plus difficile, d'où un grand nombre d'erreurs. Une amélioration de la détection bruit/parole (DBP) est donc nécessaire afin d'améliorer les performances du système de reconnaissance en environnement bruyant. Dans des travaux précédents, pour augmenter la robustesse de la reconnaissance dans les SVIs, un automate adaptatif a été développé au CNET et a donné de bonnes performances dans le cas de données RTC [4]. Cependant, le comportement de cet algorithme se dégrade avec des données GSM. Il s'agit, ici, d'améliorer la robustesse de la DBP dans le cas du réseau GSM, tout en restant efficace dans le cas du réseau RTC.

Dans ce papier, on commence par décrire, dans la section 2, le module de détection bruit/parole. Dans cette même section, on rappelle la procédure d'évaluation de la DBP et on l'illustre par un exemple sur des données GSM que l'on décrit. Dans la section 3, on introduit un nouveau critère qui tend à améliorer la robustesse de la détection bruit/parole. L'algorithme qui en résulte est ensuite évalué sur le corpus de parole à travers le GSM ainsi que sur un corpus d'exploitation en environnement RTC. Les résultats d'évaluation sont regroupés en section 4.

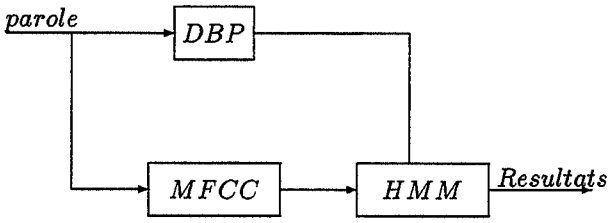


Figure 1: Schéma du système de reconnaissance. MFCC est le module d'analyse acoustique et DBP est celui de détection B/P.

2. DÉTECTION BRUIT PAROLE

Afin de déterminer les parties du signal de parole à reconnaître, on place en amont du système de reconnaissance un module de détection bruit/parole. Le sytème global est représenté par la figure 1.

Le rôle de ce module consiste à délimiter, ne serait-ce que grossièrement, les segments de parole. Des modèles de silence "début" et silence "fin" introduits dans l'algorithme de reconnaissance permettent d'affiner ces segments et rendent ainsi la détection plus précise.

Dans notre cas, le module de détection bruit/parole utilisé repose sur un automate dont les changements d'état sont conditionnés par l'énergie des trames et les durées de séjour dans les états.

2.1. *Description du fonctionnement de la DBP actuelle*

Le système de détection bruit/parole actuel utilise un automate à 5 états (*silence, parole, présomption de parole, plosive ou silence, reprise possible de parole*). Le passage d'un état à l'autre est conditionné par un seuil sur l'énergie

du signal et par des contraintes de durée (e.g. voyelle ou plosive). La transition entre les états permet de délimiter les différents segments. Les frontières de la parole dans le signal sont déterminées par le passage par l'état *parole*.

Dans le cas d'un détecteur adaptatif, l'énergie intervient en terme d'estimation du rapport signal à bruit dans le signal de parole observé. Ce rapport est ensuite comparé à un seuil donné afin de décider de la présence ou non d'un segment à détecter. Une description précise du fonctionnement et de l'algorithme de la DBP se trouve dans [4].

2.2. *Evaluation de la DBP*

L'évaluation du système de DBP se fait à deux niveaux :

1- On commence par une évaluation intrinsèque de la segmentation en comparant les segments détectés par l'automate à des segments de référence délimités manuellement. Parmi les erreurs comptabilisées, certaines sont *rejetables* par une procédure de rejet dans le système de reconnaissance, comme les fausses détections; et d'autres sont *définitives* comme les non-détections.

2- Ensuite, le but final étant la reconnaissance, on effectue une évaluation globale en considérant tous les modules du système de reconnaissance. De cette façon, on tient compte des erreurs commises lors de la détection et qui peuvent être rattrapées par un autre module (rejet). Par exemple, un bruit détecté peut-être correctement rejeté (par le module de rejet), ce qui rattrape l'erreur de détection. La procédure d'évaluation globale (détaillée dans [4]) est basée sur la comparaison des marqueurs segmentés et étiquetés manuellement et ceux obtenus par l'algorithme de détection et leur étiquetage par un modèle de reconnaissance.

2.3. *Exemples de résultats d'évaluation de l'automate actuel*

Description du corpus de test L'automate a été évalué sur un corpus de parole GSM de 51 mots distincts. Chaque communication correspond à un locuteur différent. L'enregistrement s'est fait de façon continue, pour chaque locuteur, sans tenir compte des pauses entre les mots du vocabulaire ; c'est-à-dire que toute la communication est enregistrée. Ceci justifie l'appellation "communications continues" que l'on attribue aux enregistrements de ce corpus. Les signaux de parole correspondant contiennent alors tous les silences et les bruits survenus lors des appels.

Après segmentation et étiquetage manuels de ces "communications continues", on obtient un corpus contenant 35995 segments dont 64% sont des mots du vocabulaire, 7% hors vocabulaire et 29% de bruits (cf [5]). Pour les différents tests, on utilise un modèle par mot et des marqueurs de référence obtenus par segmentation et étiquetage manuels.

Evaluation de la DBP initiale L'évaluation est exprimée en termes de pourcentage d'erreurs de reconnaissance et résumée séparément pour chaque environnement d'appel dans la figure 2. On distingue les erreurs de substitution, les fausses acceptations et le rejet à tort; ce dernier englobe les non-détections (erreurs dues au module de DBP) et le faux rejet dû au module de reconnaissance.

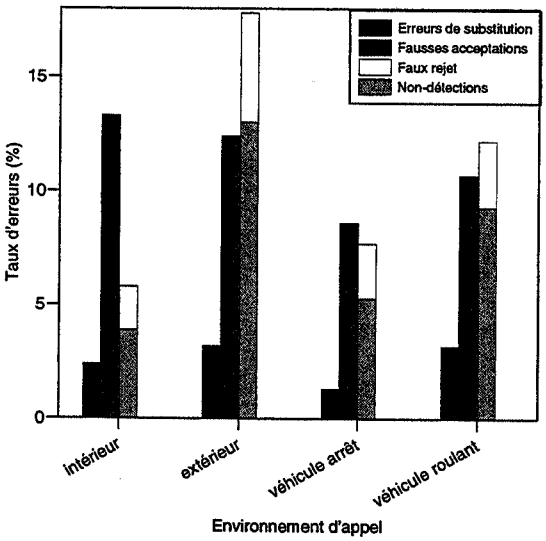


Figure 2: Evaluation selon l'environnement. Les pourcentages des erreurs de substitution, de fausses acceptations et de rejet à tort sont calculés sur le système global, en considérant chaque environnement d'appel. On distingue les non-détections (rejet dû à la DBP) et le faux rejet dû à la reconnaissance même.

On remarque que les environnements difficiles (les plus bruités : appels de véhicules roulants et surtout de l'extérieur) sont ceux qui dégradent le plus les performances. Les taux de non-détection sont notamment plus élevés dans ces environnements.

On peut aussi remarquer que les taux d'erreurs les plus élevés correspondent aux non-détections. La DBP n'est donc pas suffisamment sensible pour détecter des segments de parole noyés dans le bruit. Il est ainsi nécessaire de tenir compte des variations du signal de parole dans le module de DBP afin d'augmenter la robustesse de la reconnaissance en milieux bruités.

3. CRITÈRE STATISTIQUE POUR L'AUTOMATE DE DBP

On reprend le même automate à 5 états décrit dans 2.1. En revanche l'algorithme de détection B/P qui gère le passage d'un état à l'autre est modifié : on remplace le critère du rapport signal à bruit par un critère statistique.

Dans le cas présent, on cherche à attribuer une trame donnée du signal observé à une classe : silence (bruit) ou parole. Pour cela, on considère l'énergie de la trame et on teste l'hypothèse que ce soit une trame de silence. On choisit pour ce test une loi normale de paramètres : (μ, σ^2).

La décision sera prise en comparant le rapport critique $r(x) = \frac{x-\mu}{\sigma}$ à un seuil [6, 7] et en tenant compte des durées de séjour dans chaque état.

Par ailleurs, compte tenu de la variabilité du signal de parole (due à l'environnement), la moyenne et la variance du bruit sont mises à jour, de manière adaptative, à chaque passage par l'état de silence.

4. RÉSULTATS D'ÉVALUATION

4.1. *Evaluation en environnement GSM*

L'algorithme obtenu a été évalué sur le corpus de parole GSM évoqué ci-dessus en utilisant le même modèle par mot et des marqueurs de référence obtenus par segmentation et étiquetage manuels (cf [5]).

Les résultats de l'évaluation de la segmentation sont regroupés dans la figure 3. Les courbes sont obtenues pour différentes valeurs du seuil de fonctionnement des deux algorithmes.

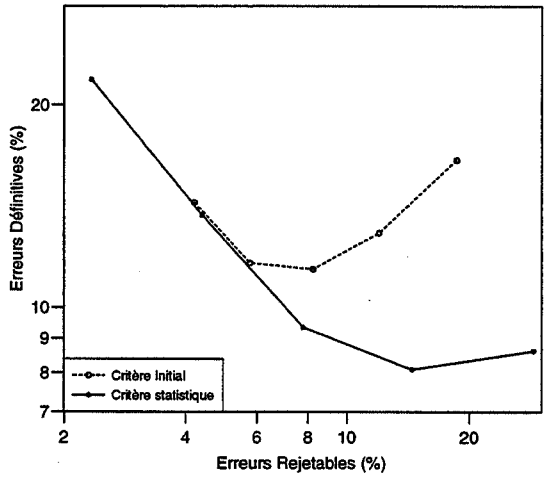


Figure 3: Evaluation de la segmentation en environnement GSM. Les différents points de chaque courbe sont obtenus en faisant varier le seuil de décision.

On remarque alors, pour un même taux d'erreurs rejetables, une diminution significative des erreurs définitives par rapport à la DBP initiale.

Cette amélioration est confirmée par les résultats de l'évaluation globale résumés dans la figure 4 où l'on représente le taux d'erreurs graves (erreurs de substitution et fausses acceptations) en fonction des erreurs moins graves (rejet à tort). Ils sont obtenus en appliquant l'ancien algorithme puis le nouveau sur le signal original. Les courbes sont obtenues en faisant varier le poids du rejet.

On obtient alors une amélioration sensible des résultats par rapport à la DBP initiale.

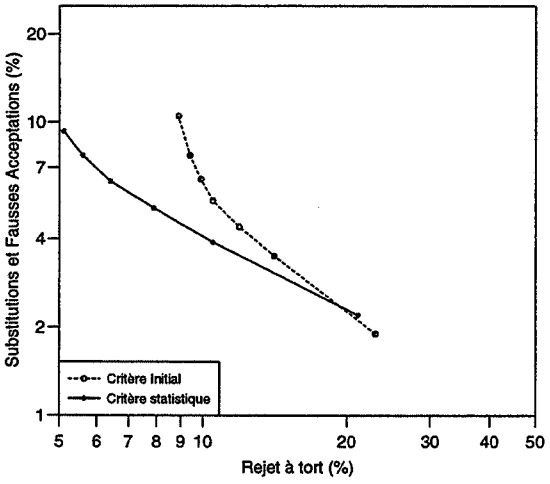


Figure 4: Taux d'erreurs graves en GSM. Les courbes sont obtenues en faisant varier le rejet.

4.2. *Evaluation en environnement RTC*

D'un autre côté, et afin de vérifier la compatibilité des résultats, on teste le nouvel algorithme sur le corpus des *Baladins* qui est enregistré à partir du réseau RTC. C'est une base de données d'exploitation collectée en utilisant un serveur vocal interactif : *"Les Baladins"*. Le vocabulaire de ce corpus contient 26 mots. 30500 segments ont été marqués manuellement dans le corpus. Parmi ces segments, 69 % sont des mots du vocabulaire, 11 % des mots hors vocabulaire et 20% de bruits divers (cf [5] pour plus de détails).

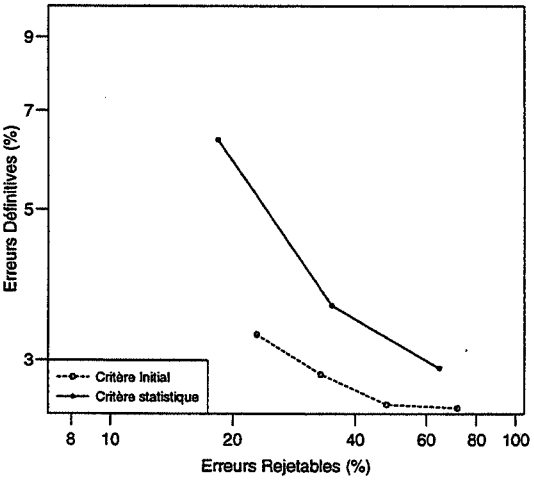


Figure 5: Evaluation de la segmentation en utilisant l'ancien critère et le nouveau avec des données RTC. Les courbes sont obtenuss en faisant varier le seuil de décision.

On fait les mêmes tests que dans le cas du GSM et on obtient les résultats résumés dans les figures 5 et 6.

On remarque que pour les données RTC le nouveau détecteur B/P dégrade la segmentation. Cette dégradation peut-être due au fait que cet algorithme, étant plus sensible aux variations du signal, détecte plus de segments augmentant ainsi le nombre de détections correctes mais aussi le risque d'insertions. Parmi ces insertions on compte les insertions de parole hors vocabulaire, ce qui augmente le taux d'erreurs qui peuvent être rejetées par la suite. En effet, en utilisant de bons modèles de rejet, l'évaluation globale permet de rattraper un grand nombre de mots hors vocabulaire détectés par le nouvel algorithme. Ceci est clairement visible sur la figure 6, qui résume les résultats de l'évaluation globale avec les deux algorithmes de DBP.

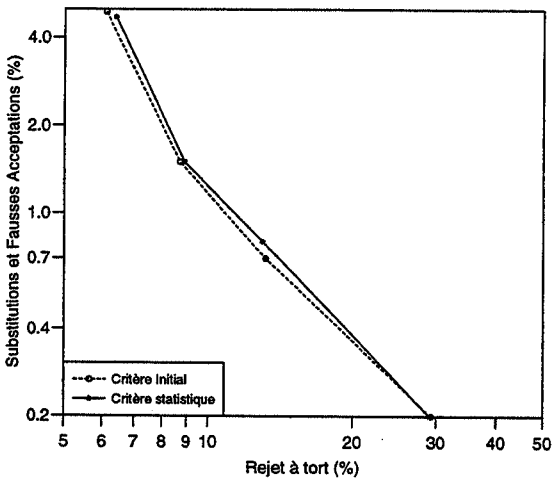


Figure 6: Taux d'erreurs graves en RTC. Les résultats sont représentés pour l'ancien critère et le nouveau en faisant varier le rejet.

L'écart entre les deux algorithmes est alors réduit et le nouveau critère de l'algorithme de DBP, efficace en GSM, reste compatible en RTC.

5. CONCLUSIONS

Pour améliorer le fonctionnement du module de détection bruit/parole, on a introduit un critère statistique pour remplacer le critère initial basé sur une estimation du rapport signal à bruit dans la communication traitée. On obtient ainsi un nouvel algorithme de DBP basé sur l'estimation des statistiques du bruit dans le signal de parole observé. Une évaluation globale de tout le système de reconnaissance montre une amélioration des résultats par rapport à l'algorithme actuellement utilisé. On montre que cet algorithme est robuste aux variations du signal de parole en milieu bruité. Il permet, par exemple dans le cas du GSM, une réduction de 30% du taux d'erreurs graves pour un même taux de rejet à tort égal à 10%. Par ailleurs, l'évaluation globale sur un corpus RTC donne des résultats équivalent à ceux de l'ancien algorithme. On remarque notamment que beaucoup d'erreurs de segmentation ont été rattrapées par le module de rejet du système global. Ceci prouve l'importance d'un bon choix de modèles de rejet que l'on doit intégrer dans les systèmes de reconnaissance vocale.

BIBLIOGRAPHIE

[1] L. Karray and L. Mauuary, " Improving Speech Detection for Wireless Speech Recognition", Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU97), pp. 428-435, Santa Barbara (CA), Décembre 1997.

[2] J.C. Junqua *et al.* , " A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE Trans. on Speech Audio Processing, pp. 406-412, Vol. 2, No. 3, July 1994.

[3] H. Agaiby and T.J. Moir, " Knowing the Wheat from the Weeds in Noisy Speech," Proc. Eurospeech97, pp. 1119-1122, Rhodes, Greece, September 1997.

[4] L. Mauuary and J. Monné , "Speech/non-Speech Detection for Voice Responses Systems," Proc. Eurospeech93, pp. 1097-1100, Berlin, Septembre 1993.

[5] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin and K. Bartkova "Towards Improving ASR Robustness for PSN & GSM Telephone Applications," to appear in Speech Communication Journal.

[6] D.H. Sanders, A.F. Murph and R.J. Eng, "Les statistiques , une approche nouvelle," McGraw - Hill, Canada, 1984.

[7] G. Saporta, "Probabilités, Analyse des données et statistique," Editions Technip, Paris, 1990.

Utilisation des Modèles de Markov Cachés pour le Débruitage

Driss MATROUF et Jean-Luc GAUVAIN

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
Tél.: 01 69 85 80 67 - Fax: 01 67 85 80 88
e-mail: driss@limsi.fr - http://www.lmsi.fr/TLP

RÉSUMÉ

In this paper we address the problem of enhancing speech which has been degraded by additive noise. As proposed by Ephraim et al., autoregressive hidden Markov models (AR-HMM) for the clean speech and an autoregressive Gaussian for the noise are used. The filter applied to a given frame of noisy speech is estimated using the noise model and the autoregressive Gaussian having the highest *a posteriori* probability given the decoded state sequence. The success of this technique is highly dependent on accurate estimation of the best state sequence. A new strategy combining the use of cepstral-based HMMs, autoregressive HMMs, and a model combination technique, is proposed. The intelligibility of the enhanced speech is indirectly assessed via speech recognition, by comparing performance on noisy speech with compensated models to performance on the enhanced speech with clean-speech models. The results on enhanced speech are as good as our best results obtained with noise compensated models.

1. INTRODUCTION

Parmi les techniques de débruitage qui ont été les plus étudiées pendant les deux dernières décennies, on peut citer la technique de la soustraction spectrale [8, 9]. La soustraction spectrale, sous toutes ses formes, appliquée à un signal de parole bruité, afin de produire le signal propre correspondant, ne tire pas profit de l'information linguistique portée par ce signal. En fait, la plupart des implantations de la soustraction spectrale sont appliquées à un signal de parole exactement de la même manière qu'à n'importe quel autre signal bruité. Cette insuffisance d'information *a priori* dans le processus de débruitage engendre d'importantes distorsions dans les signaux restaurés.

Une tentative introduisant de l'information *a priori* dans le processus de débruitage a été proposée par J.S. Lim et A.V. Oppenheim [11]. Ils ont proposé une technique qui consiste à représenter le signal de parole par une succession de modèles AR : le signal est segmenté en une suite de zones stables (stationnaires), chaque zone est modélisée par un modèle AR. Les modèles AR et le signal propre sont tous les deux estimés en utilisant le signal bruité et la densité spectrale du bruit, supposée connue *a priori*. Pour cela, l'approche du MAP (*Maximum a Posteriori*) a été utilisée. La maximisation d'une fonction de vraisemblance appropriée est réalisée de manière itérative : 1) par rapport aux paramètres des modèles AR en supposant que le signal dont on dispose est propre, 2) par rapport au signal propre en utilisant les modèles AR et l'estimation de la densité spectrale du bruit. Il est évident que cette procédure itérative possède beaucoup plus de variables inconnues (signal propre et modèles AR) que de variables connues (signal bruité). Ceci conduit à des estimateurs, du signal et des modèles, avec de grandes variances. Pour pallier ce problème, Ephraim et al. [7] proposent d'estimer les modèles AR représentant le signal propre à partir de signaux d'apprentissage propres au lieu de les estimer directement sur le signal bruité.

Afin de déterminer les trames de parole propre maximisant la densité de probabilité *a posteriori*, Ephraim et al. [7] utilisent une procédure itérative fondée sur l'algorithme EM. Ce processus itératif est très dépendant de l'initialisation. Lorsque l'initialisation est mauvaise, ce processus converge vers un maximum local qui peut être très loin de la solution optimale. Cela est toujours le cas avec des signaux très bruités. Logan et Robinson [10] proposent d'utiliser une technique de combinaison de modèles dans le cadre des MMC autorégressifs (MMCAR) permettant de mieux initialiser ce processus. L'initialisation est obtenue par décodage du signal bruité en utilisant le système de reconnaissance fondé sur des MMCAR adaptés au bruit de test. Cependant, il est bien connu qu'un système de reconnaissance utilisant des MMC fondés sur le cepstre et ses dérivées est plus efficace qu'un système de reconnaissance utilisant des MMCAR, surtout dans le cas d'applications avec de grands vocabulaires. Dans cet article, nous proposons une extension de cette dernière approche en utilisant un système de reconnaissance fondé sur le cepstre et ses dérivées pour l'initialisation.

Nous proposons d'utiliser deux ensembles de modèles acoustiques : le premier est un MMC à densités continues (MMCDC) fondé sur le cepstre utilisé pour l'initialisation du processus itératif ; le second est un MMCAR utilisé pour estimer la succession de filtres à appliquer aux trames constituant le signal bruité. Pour une trame du signal bruité, nous cherchons la gaussienne cepstrale qui a la plus grande probabilité *a posteriori*. Cette recherche est faite par décodage du signal bruité en utilisant les modèles cepstraux adaptés au bruit de test. Le décodage fournit l'alignement trame/gaussienne-cepstrale, où chaque gaussienne cepstrale correspond à une gaussienne autoré-

gressive dans les MMCAR. Le filtre optimal est donc estimé en utilisant cette gaussienne autorégressive et la gaussienne autorégressive correspondant au bruit. Les MMC cepstraux et les MMCAR sont entraînés de façon à avoir une bijection entre les deux ensembles de modèles au niveau de la gaussienne : à une gaussienne dans les MMC cepstraux correspond une et une seule gaussienne dans les MMCAR. Pour cela, nous estimons en premier les MMC cepstraux et nous utilisons les statistiques correspondant à la dernière itération de l'algorithme EM pour estimer les paramètres des MMCAR.

Dans les sections suivantes nous décrivons la mise en œuvre de cette approche de débruitage. Des expériences avec différents types de bruits sont présentées afin de montrer l'apport de cette approche à l'amélioration de la qualité des signaux de parole bruités. Des expériences de reconnaissance avec les signaux débruités seront présentées .

2. LE PROCESSUS DE DÉBRUITAGE

Considérons, dans un premier temps, le cas d'une observation générée par une seule gaussienne. Soit $\mathbf{y}$ une trame du signal de parole bruité ($\mathbf{y} \in \mathbf{R}^K$) et $f_{\lambda_\mathbf{x}}$ la fonction de densité de probabilité (fdp) de la trame propre $\mathbf{x}$ correspondant à la trame bruitée $\mathbf{y}$. En supposant que $\mathbf{x}$ est générée par un processus autorégressif, sa fdp $f_{\lambda_\mathbf{x}}(\mathbf{x})$ est définie comme suit :

$$f_{\lambda_\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}} |S_\mathbf{x}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}' S_\mathbf{x}^{-1} \mathbf{x})\} \quad (1)$$

où $S_\mathbf{x}$ désigne la matrice d'autocorrélation : $S_\mathbf{x} = \sigma_\mathbf{x}^2 (A_\mathbf{x}' A_\mathbf{x})^{-1}$, où $\sigma_\mathbf{x}^2$ est la variance de l'excitation du modèle AR, $A_\mathbf{x}$ est la matrice $K \times K$ triangulaire inférieure dont les $p+1$ premiers éléments de la première colonne sont les coefficients du processus AR : $a_{i,0 \leq i \leq p}$ où $a_0 = 1$.

Supposons également que la fonction de densité de probabilité du bruit additif peut être modélisée par une gaussienne autorégressive, soit $f_{\lambda_\mathbf{n}}$, le problème du débruitage consiste à estimer la trame propre $\mathbf{x}$ en utilisant $f_{\lambda_\mathbf{x}}$, $f_{\lambda_\mathbf{n}}$ et $\mathbf{y}$. Le critère du Maximum *A Posteriori* (MAP) peut être utilisé pour réaliser cette estimation :

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \log\{h(\mathbf{x}, \mathbf{y})\} \quad (2)$$

où $h(\mathbf{x}, \mathbf{y})$ désigne la fdp conjointe de $\mathbf{x}$ et $\mathbf{y}$. Le bruit étant additif et indépendant du signal, nous avons :

$$h(\mathbf{x}, \mathbf{y}) = f_{\lambda_\mathbf{x}}(\mathbf{x}) f_{\lambda_\mathbf{n}}(\mathbf{y} - \mathbf{x}). \quad (3)$$

Cette maximisation aboutit au filtre de Wiener bien connu dans la littérature du traitement de signal. La transformée de Fourier de l'estimation du signal propre est :

$$\hat{X}(\theta) = \frac{\Gamma_\mathbf{x}(\theta)}{\Gamma_\mathbf{x}(\theta) + \Gamma_\mathbf{n}(\theta)} Y(\theta) \quad (4)$$

où $Y(\theta)$ est la transformée de Fourier de la trame de parole bruitée, $\Gamma_\mathbf{x}(\theta)$ et $\Gamma_\mathbf{n}(\theta)$ sont les densités spectrales de puissance correspondantes aux deux processus AR $\mathbf{x}$ et $\mathbf{n}$. Les densités spectrales de puissance sont déterminées comme suit :

$$\Gamma_\mathbf{x}(\theta) = \frac{\sigma_\mathbf{x}^2}{|\Psi_\mathbf{x}(\theta)|^2}, \quad (5)$$

$$\Gamma_\mathbf{n}(\theta) = \frac{\sigma_\mathbf{n}^2}{|\Psi_\mathbf{n}(\theta)|^2} \quad (6)$$

où $\Psi_\mathbf{x}(\theta)$ et $\Psi_\mathbf{n}(\theta)$ désignent, respectivement, les transformées de Fourier des coefficients de prédiction associés à la trame de la parole propre et des coefficients de prédiction associés au bruit.

Considérons, maintenant, le cas général d'une observation générée par un MMCAR. Soit $\mathbf{y} = \mathbf{y}_{t,t=1,\ldots,T} / \mathbf{y}_t \in \mathbf{R}^k$ une séquence de trames correspondant à une phrase bruitée (T est le nombre de trames constituant le signal correspondant à la phrase en question). L'estimateur MAP des trames correspondant au signal propre est obtenu itérativement en utilisant l'algorithme EM. À chaque itération k, la transformée de Fourier $\hat{X}_t^k$, $\{t = 1, \ldots, T\}$ de l'estimation de la trame propre $\hat{\mathbf{x}}_t(k)$ est obtenue par filtrage de la trame bruitée comme suit :

$$\hat{X}_t^{k+1}(\theta) = \left[\sum_{\beta,\gamma} p_t(\beta, \gamma | \hat{\mathbf{x}}(k)) H_{\gamma|\beta}^{-1} \right]^{-1} Y_t(\theta), \quad (7)$$

où $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$ est la probabilité que la trame $\hat{\mathbf{x}}_t(k)$ soit générée à l'instant t par la gaussienne γ de l'état β, sachant que $\hat{\mathbf{x}}(k)$ est générée par le MMCAR. $H_{\gamma|\beta}$ est le filtre de Wiener associé à la gaussienne autorégressive γ de l'état β et à la gaussienne autorégressive du bruit (Equation 4).

Le succès de cette procédure d'estimation est très dépendant de la qualité d'estimation des probabilités *a posteriori* $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$. Ces probabilités sont estimées en utilisant la procédure "avant-arrière". L'estimation de ces probabilités en utilisant les modèles acoustiques entraînés sur de la parole propre donne lieu à une procédure itérative convergeant vers un maximum local qui est loin d'être la solution optimale recherchée, surtout pour des rapport signal sur bruit (RSB) faibles. Pour obtenir une bonne estimation des probabilités *a posteriori*, nous décodons le signal bruité $\mathbf{y}$ en utilisant des modèles acoustiques obtenus par adaptation au bruit de test des modèles acoustiques cepstraux retenus pour la parole propre (MMCDC), c'est-à-dire les meilleurs modèles disponibles pour la parole bruitée.

Plusieurs techniques ont été proposées pour estimer les modèles représentant la parole bruitée à partir des modèles représentant la parole propre et du modèle du bruit. Parmi ces techniques, on peut citer la Combinaison Parallèle de Modèles (CPM) et l'approche par génération de données [2]. Dans ce travail, nous utilisons une technique fondée sur la composition directe sur les données (CPD : Combinaison Parallèle de Données) [1]. Elle consiste à réutiliser pour chaque gaussienne d'un état donné, d'un modèle donné, les vecteurs cepstraux d'apprentissage qui ont permis son estimation. Comme pour la CPM, on suppose que le bruit n'affecte pas la distribution des trames entre les gaussiennes d'un état donné. Afin d'améliorer d'avantage l'estimation des probabilités *a posteriori*, il est possible, en plus de la CPD, d'utiliser des techniques d'adaptation aveugles comme la MLLR [5]. Dans ce travail, seule la CPD est utilisée pour l'adaptation, les composantes principales du système de débruitage proposé dans cet article sont présentées dans la Figure 1.

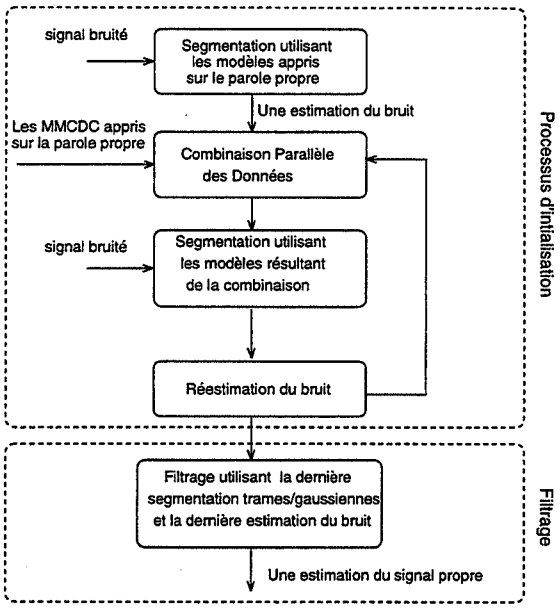


Figure 1: Le processus de débruitage utilisant le filtrage de Wiener dépendant de l'état dans les MMCAR. Les MM-CDC cepstraux et la CPD sont utilisés pour avoir une bonne estimation des probabilités *a poteriori*.

3. RÉSULTATS EXPÉRIMENTAUX

Afin d'évaluer cette technique de débruitage, nous utilisons un système entraîné sur 22148 phrases prononcées par 460 locuteurs, provenant du corpus MASK [4]. Les phrases prononcées par 450 locuteurs ont été utilisées pour l'apprentissage du système et les phrases prononcées par les 10 locuteurs restants ont été utilisées pour le test. Ces données ont été enregistrées avec un microphone de tête résultant en un RSB d'environ 35dB. Le signal de parole a subi un filtrage passe-bas avec une fréquence de coupure égale à 8kHz et a été échantillonné à 16kHz.

Pour les deux systèmes (cepstral et autorégressif) les vecteurs paramétriques sont estimés tous les 10ms sur une fenêtre de 30ms. Pour le système cepstral, les vecteurs paramétriques caractérisant une trame sont composés de 13 coefficients cepstraux auxquels sont ajoutées leurs dérivées premières et secondes (vecteur à 39 paramètres). La soustraction du cepstre moyen est réalisée sur chaque phrase. L'ordre des gaussiennes autorégressives est fixé à 16 pour les MMCAR et pour le modèle autorégressif correspondant au bruit. 608 phonèmes dépendant du contexte sont modélisés, chaque phonème dépendant du contexte est un MMCDC à trois états avec un mélange de gaussiennes comme densité d'observation (typiquement 20 gaussiennes).

L'algorithme de débruitage a été appliqué à des signaux de parole dégradés par un bruit additif. Dans cet article trois bruits extraits de la base de données NOISEX-92 [12] ont été utilisés : le bruit blanc, le bruit d'hélicoptère (Lynx) et le bruit d'avion à réaction (F16 Jet). Le signal restauré est obtenu en utilisant la technique standard d'*overlap-add* sur les segments résultant du filtrage, suivi du fenêtrage par Hanning.

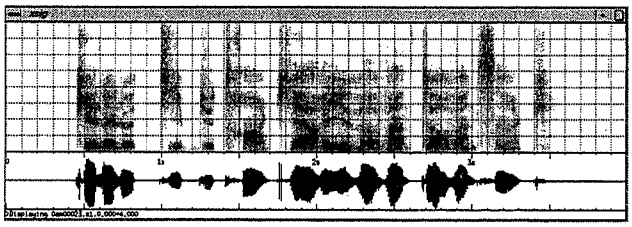


Figure 2 : *Spectrogramme du signal propre : "quel est le type du train qui arrive à 20 heures 25."*

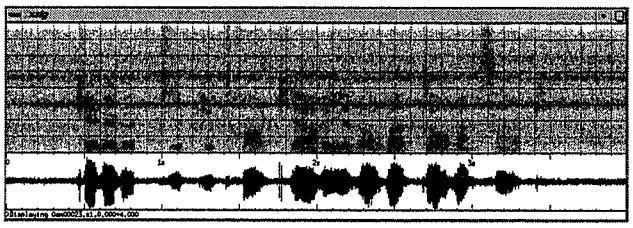


Figure 3 : *Signal bruité (RSB=5.7dB) généré en additionnant le bruit F16 Jet au signal de la Figure 2.*

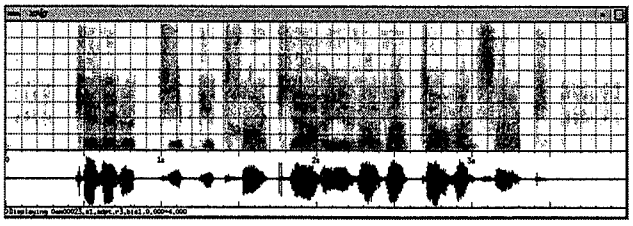


Figure 4 : *Signal obtenu par débruitage du signal de la Figure Figure 3. Le débruitage utilise le filtrage de Wiener dépendant de l'état.*

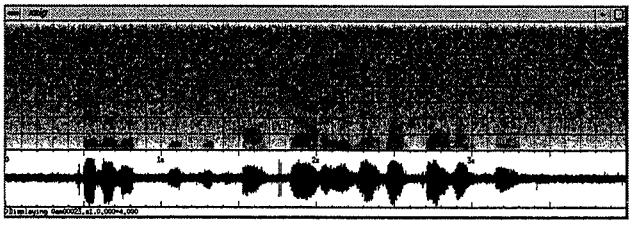


Figure 5 : *Signal bruité (RSB=1dB) généré en additionnant le bruit blanc au signal de la Figure 2.*

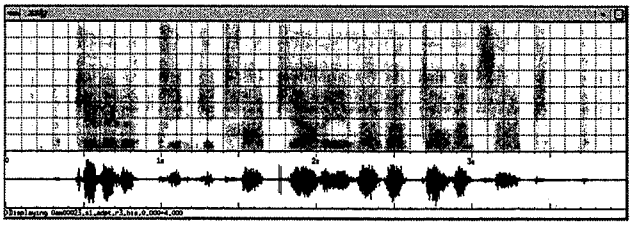


Figure 6 : *Signal obtenu par débruitage du signal de la Figure Figure 5. Le débruitage utilise le filtrage de Wiener dépendant de l'état.*

Dans la Figure 2 nous présentons le spectrogramme d'un signal propre. Les versions bruitées de ce signal sont obtenues en additionnant le bruit d'avion à réaction (F16 Jet) (Figure 3) et le bruit blanc (Figure 5). Les spectrogrammes des signaux obtenus par débruitage par application du filtrage de Wiener dépendant de l'état sont présentés dans les Figures 4 et 6 respectivement. Visuellement, une diminution considérable du bruit est observée par débruitage. En écoutant les signaux restaurés nous constatons une amélioration significative de la qualité avec de très faibles distorsions et l'absence du bruit musical.

La reconnaissance avec les signaux débruités

Afin d'évaluer objectivement cette procédure de débruitage, des expériences de reconnaissance avec des signaux débruités ont été réalisées. Ces expériences sont présentées dans le but de montrer qu'il n'y a pas de perte d'intelligibilité du point de vue du système de reconnaissance et non dans le but d'améliorer le taux de reconnaissance. En effet, la théorie prédit qu'aucun gain significatif en terme de taux de reconnaissance n'est à espérer par débruitage par rapport aux techniques de bruitage si les modèles mis en jeu sont de même qualité. L'objectif est l'amélioration de la qualité du signal sans perte d'intelligibilité.

Les taux d'erreur sur les mots dans différentes configurations sont présentés dans la table 1. En plus du filtrage de Wiener donné par l'équation 4 (filtre 1), nous avons utilisé le filtre résultant de l'application de la racine carrée au filtre de Wiener (filtre 2).

	% d'erreur sur les mots		
Configuration de test	F16	Lynx	Blanc
Propre	5.9	5.9	5.9
Bruitée	55.4	60.7	79.9
Compensation (CPD)	13.6	21.4	15.2
Débruitage (filtre 1)	13.9	21.7	15.2
Débruitage (filtre 2)	12.2	20.3	14.5

Table 1 *Taux d'erreur moyens sur les mots dans différentes configurations de test. La colonne 1 correspond au bruit de F16 Jet (RSB=6.4), la colonne 2 correspond au bruit de Lynx (RSB=5.5dB) et la colonne 3 correspond au bruit blanc (RSB=1dB).*

La table 1 montre une très grande détérioration des performances lorsque le système est entraîné sur des signaux propres et testé sur des signaux bruités. Par exemple, lorsque le système est entraîné sur des signaux propres et testé sur des signaux bruités (bruit blanc : RSB=1dB), le taux d'erreur sur les mots augmente de 5.9% à 79.9%. La compensation du bruit de test en utilisant la CPD apporte un gain significatif (comparer la ligne 2 et 3). En utilisant des modèles acoustiques entraînés sur des signaux propres pour décoder les signaux restaurés par filtrage de Wiener dépendant de l'état, nous obtenons des résultats similaires à ceux obtenus en utilisant la CPD. En utilisant le système entraînés sur des signaux propres pour décoder les signaux restaurés par filtrage utilisant le filtre 2, un gain relatif de 4.6% est observé dans le cas du bruit blanc, 10.3% dans le cas du bruit F16 Jet, et 5.1% dans le cas du bruit de Lynx, par rapport aux résultats obtenus en utilisant la CPD. Ces résultats montrent que du point de vue du système de reconnaissance, la technique de débruitage proposée ne provoque pas de perte d'intelligibilité des signaux restaurés par rapport aux signaux bruités.

4. CONCLUSION

Dans cet article nous avons étudié et expérimenté une approche de débruitage fondée sur le critère du maximum *a posteriori* utilisant des MMC autorégressifs. Cette technique introduite par Ephraim et al. utilise l'algorithme EM qui nécessite dans ce cas une bonne initialisation pour fonctionner correctement. Nous avons proposé une nouvelle stratégie d'initialisation combinant l'utilisation de MMCAR, l'utilisation de MMCDC fondés sur le cepstre et ses dérivées et l'utilisation d'une technique de compensation de bruits que nous avons présenté dans des travaux ultérieurs [6]. Cette stratégie d'initialisation rend ce processus de débruitage très efficace même avec des rapports signal sur bruit faibles. Les expériences ont été réalisées en utilisant trois types de bruits provenant de la base de données NOISEX. Nous avons constaté une augmentation considérable de la qualité des signaux débruités, avec peu de distorsion et sans bruit musical. Des tests de reconnaissance avec les signaux débruités montrent qu'il n'y a pas de diminution de l'intelligibilité du signal débruité par rapport au signal bruité du point de vue du système de reconnaissance.

REFERENCES

[1] J.L. Gauvain, L. Lamel, G.Adda, D.Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[2] M.J.F. Gales, S.J. Young, "A fast and flexible implementation of parallel model combination", *ICASSP-95*, pp. 133-136.

[3] A. Acero, R.M. Stern, "Environmental Robustness in Automatic Speech Recognition," *IEEE Acoustics, Speech & Signal Processing*, pp. 849-852. April 1990.

[4] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*, Madrid.

[5] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech & Language*, pp. 171-185, 1995.

[6] D.Matrouf, J.L. Gauvain, "Model compensation for noises in test and training data," *ICASSP-97*.

[7] Y. Ephraim, D. Malah, B.H Juang, "On the Application of Hidden Markov Models for Enhancing Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37. NO. 12. December 1989.

[8] S.F Boll, " Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE*, pp. 113-120, 1979.

[9] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", *ICASSP-79*, pp. 208-211.

[10] B. T. Logan, A. J. Robinson, "Enhancement and Recognition of Noisy Speech within an AutoRegressive HMM Framework Using Noise Estimates from The Noisy Signal," *ICASSP-97*.

[11] J.S. Lim, A.V. Oppeenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 26, pp. 197-210, June 1978.

[12] A.P. Varga, H.J.M Steeneken, M. Tomlinson, D. Jones,"The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *In Technical Report, DRA Speech Research Unit*, 1992.

CLASSIFICATION PAROLE/MUSIQUE DE SIGNAUX SONORES PARAMETRISES PAR LES COMPOSANTES PRINCIPALES DU SPECTRE

Mouhamadou SECK, Frédéric BIMBOT et Bernard DELYON

IRISA - SIGMA2 / INRIA & C.N.R.S.
Campus Universitaire de Beaulieu
35042 RENNES cedex, FRANCE

e-mail : mseck@irisa.fr, bimbot@irisa.fr, delyon@irisa.fr

ABSTRACT

We present a technique for the classification of sound segments into speech or music. Each frame of signal is preprocessed by extracting principal components of the logarithm of the spectrum magnitude. A multidimensional gaussian classifier is used for frame recognition. Decision for the segment is taken by majority vote on all its frames. An acoustic feature selection method, using a cross-validation data set, selects most of the the first principal components as containing more discriminant information. With approximatively 25 principal components, we reach a performance level around 95% segments classified correctly. Principal components seem very similar to cepstral coefficients, but cepstral coefficients yield a lower recognition rate.

1. INTRODUCTION

L'indéxation de bandes sonores constitue un domaine auquel s'intéresse un nombre croissant de chercheurs. Cet intérêt s'explique par les perspectives d'application dans des secteurs en pleine évolution comme le traitement du son, la gestion de banques de données audiovisuelles, le multimédia...

Une des tâches à résoudre dans ce contexte est la classification parole/musique de segments sonores. Dans ce but, Saunders [Sau96] utilise une paramétrisation acoustique basée sur la fréquence des passages par zéro du signal, décrite par Kedem [Ked86] comme un outil de discrimination rapide de signaux. Sheirer et Slaney [Sch97] proposent un système de classification qui intègre une gamme de paramètres pouvant se prêter à une interprétation en terme de discrimination parole/musique. L'approche que nous proposons fait référence à des outils simples de statistique et d'analyse de données, en faisant abstraction d'hypothèses *a priori* sur la nature des informations discriminantes entre la parole et la musique.

La section 2 de cet article est consacrée à la description de l'algorithme mis en œuvre. La section 3 traite de la sélection des paramètres acoustiques et la section 4 présente le protocole expérimental utilisé. Les résultats expérimentaux sont présentés dans la section 5. Dans la dernière section, nous comparons deux méthodes de sélection des paramètres acoustiques, puis nous étudions leurs relations avec les coefficients cepstraux.

2. PRINCIPE DU SYSTÈME

2.1. Fonctionnement

Le système de classification que nous utilisons suit un schéma classique (figure 1). Le segment de signal est d'abord découpé en trames régulières. Chacune des trames est ensuite représentée par un vecteur (de paramètres) acoustique(s) (module 1). Dans un premier temps, les vecteurs acoustiques de chaque classe sont modélisés à l'aide de données d'apprentissage (module 2). En phase de test, on classifie un signal (de classe inconnue) en calculant d'abord les scores de ses trames par rapport à chaque classe (module 3). Ces scores sont ensuite transmis au module de décision (module 4) qui indique la classe reconnue.

Comme paramètres acoustiques, nous utilisons les composantes principales du logarithme du module du spectre (obtenu lui-même par Transformée de Fourier Discrète du signal). Nous décrivons au §3 la façon dont nous obtenons ces paramètres acoustiques. Les vecteurs acoustiques pour la parole comme pour la musique, sont modélisés par une variable gaussiennne (à covariance pleine). Le score d'une trame par rapport à une classe est ici égal à la vraisemblance du vecteur acoustique de la trame pour le modèle de la classe. En ce qui concerne la décision, chaque trame est d'abord allouée à une classe selon la règle du maximum de vraisemblance. Ensuite, la décision finale sur le segment entier est prise par un vote majoritaire sur les trames qui le composent.

2.2. Formalisation

De manière formelle, on note x le signal observé. Soit x_t la trame courante du signal, $1 \leq t \leq n$, où n est le nombre de trames du segment. On note y_t le vecteur acoustique de la trame t (vecteur de dimension k). Chaque modèle monogaussien est représenté par sa moyenne et sa matrice de covariance (respectivement μ_p, C_p pour la parole et μ_m, C_m pour la musique). Ensuite, on calcule les scores du vecteur acoustique y_t pour les modèles de parole et de musique par les formules :

$S_t^{(p)} = -\frac{1}{2}\log[det(C_p)] - \frac{1}{2}(y_t - \mu_p)^T C_p^{-1}(y_t - \mu_p)$

$S_t^{(m)} = -\frac{1}{2}\log[det(C_m)] - \frac{1}{2}(y_t - \mu_m)^T C_m^{-1}(y_t - \mu_m)$

qui correspondent aux log-vraisemblances[1] pour chacune

[1] A la constante additive $-\frac{k}{2}\log 2\pi$ près.

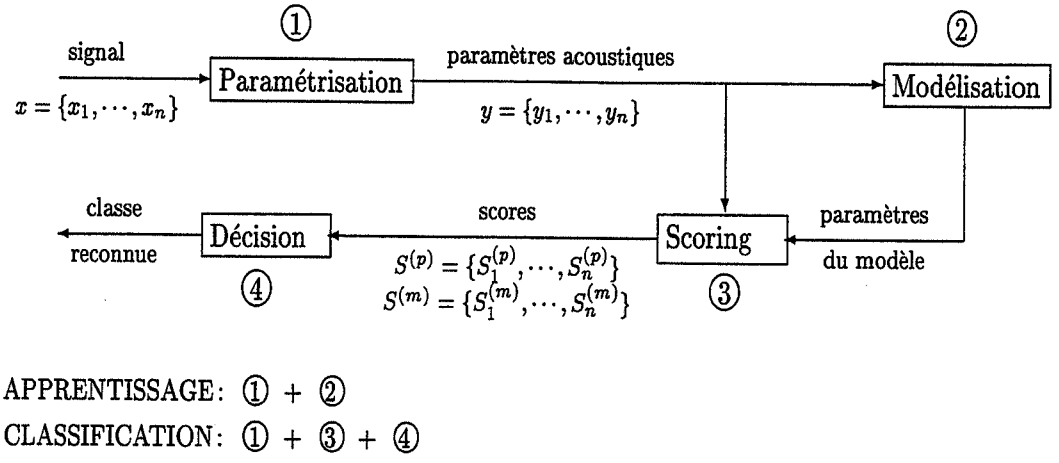


Figure 1: Schéma du système de classification.

des deux classes. Le critère du maximum *a posteriori* (MAP) est utilisé pour la décision trame-à-trame en supposant l'égalité des probabilités *a priori* [Sap90] : en d'autres termes, on décide que x_t est une trame d'un signal de parole si $S_t^{(p)} > S_t^{(m)}$. Enfin, en notant $\mathbf{1}$ la fonction indicatrice et en posant :

$$V^{(p)} = \sum_{t=1}^{n} \mathbf{1}_{[S_t^{(p)} > S_t^{(m)}]}$$

la règle du vote majoritaire consiste à décider que x est un segment de parole si : $V^{(p)} > \frac{n}{2}$.

3. Composantes Principales du spectre (CPS)

3.1. *Rappels théoriques*

L'utilisation de l'intégralité des coefficients de Fourier donne lieu à un système peu économique en calculs et susceptible d'être peu robuste. C'est pourquoi nous avons choisi d'effectuer une transformation des coefficients issus de la Transformée de Fourier, par une Analyse en Composantes Principales (ACP). Cette transformation a également pour propriété de décorréler les coefficients acoustiques.

Soit Y une variable aléatoire de dimension d et de matrice de covariance C. Considérons la décomposition en valeurs singulières de C en : $C = UDU^T$. U est la matrice unitaire dont les colonnes sont les vecteurs propres de C et D et la matrice diagonale des valeurs propres de C, ordonnées selon l'ordre décroissant. Les composantes principales de Y sont les composantes de U^TY. Les composantes principales ont pour matrice de covariance D et sont donc linéairement indépendantes. Le système des k premières composantes principales correspond au système de projection de la variable Y dans un sous-espace de dimension k, d'inertie maximale.

Dans la pratique, la matrice de covariance C est inconnue. On utilise alors une estimation de cette matrice à partir d'observations de la variable Y. Dans cet article, les composantes principales sont obtenues à partir d'une estimation empirique de la matrice de covariance. Dans la suite, nous désignerons par Composantes Principales du Spectre (CPS), les composantes principales du log-module du spectre.

3.2. *Sélection des CPS*

Le critère classique de sélection des composantes principales consiste à utiliser leur variance [Sap90]. Cependant, ce critère ne prend pas explicitement en compte le contexte d'utilisation des composantes. Une solution plus proche de notre problème consisterait à apprécier l'utilité d'une composante principale selon son pouvoir de discrimination pour la tâche à accomplir.

Dans cet article, nous comparons deux procédés de sélection des CPS. Le premier consiste à conserver les composantes de plus grande variance (sélection par la variance), le second procédé repose sur le pouvoir discriminant des composantes pour la tâche spécifique de classification parole/musique (sélection par le pouvoir discriminant), critère évalué sur un ensemble de développement. Cette approche s'apparente à celle utilisée notamment dans [Mag97] pour la reconnaissance du locuteur.

Pour un nombre k de composantes, une recherche directe de l'ensemble le plus discriminant consisterait à comparer les performances de $(2^k - 1)$ combinaisons de composantes. Le coût en temps de calcul atteind donc très rapidement des proportions non envisageables.

Nous adoptons une alternative sous-optimale de sélection par agglomération successive. Dans une étape initiale, on recherche la composante qui possède les meilleures performances. A chaque étape, on introduit la composante qui, jointe à celles déjà sélectionnées, présente les meilleures performances. Cette méthode ne donne pas forcément l'ensemble de composantes le plus discriminant, mais fournit des résultats assez satisfaisants pour une économie considérable en temps d'optimisation.

4. PROTOCOLE EXPERIMENTAL

4.1. *Les données*

La base de données est constituée d'extraits radiophoniques enregistrés sur une vingtaine de stations émettant sur la Bretagne et au sein desquels sont représentés différents :

- types de sons : parole seule, annonce publicitaire sur léger fond musical, musiques avec chansons (prédominance instrumentale), musique instrumentale;
- styles musicaux : Dance, Jazz, Rap, Reggae, Rock, Soul...;
- types de locuteurs : hommes et femmes de tous âges, locuteurs professionnels ou non;
- conditions de prise du son : personne parlant en direct à l'antenne ou en liaison téléphonique avec la station.

Les enregistrements sont effectués à la fréquence d'échantillonnage de 16 kHz, sous forme d'extraits de 4, 8 s. La base de données compte 120 extraits pour la parole et autant pour la musique.

Les données sont ensuite réparties en trois ensembles :

- des données d'apprentissage (60%), pour estimer les distributions des vecteurs acoustiques;
- des données de développement (20%), pour l'algorithme de sélection des CPS selon le pouvoir discriminant, décrit au §3.2;
- des données de test (20%), pour l'évaluation des performances.

Nous précisons que ce sont des extraits entiers qui sont alloués à un des trois ensembles, cela pour s'affranchir d'une certaine dépendance entre les trois ensembles.

4.2. *Conditions expérimentales*

Dans toutes nos expériences, nous considérons des trames d'une durée de 16 ms, soit un vecteur de dimension 256 échantillons. La partie utile de la Transformée de Fourier de la trame correspond aux $d = 129$ premiers coefficients. La classification s'effectue sur des segments de $n = 150$ trames successives, donc d'une durée de 2, 4 s (correspondant à la moitié d'un extrait de la base de données).

5. RESULTATS EXPERIMENTAUX

Nous mesurons les performances du système par la proportion de segments correctement reconnus (avec équirépartition des segments de parole et de musique dans l'ensemble de test). La figure 2 représente, pour les deux critères de sélection des CPS (la variance et le pouvoir discriminant), les taux de reconnaissance sur les données de développement en fonction du nombre de CPS choisies. La figure 3 est l'analogie de la figure 2, pour les données de test. On observe sur les données de test, qu'une performance de 95% est obtenue avec environ 25 CPS, pour les deux méthodes de sélection. Les 25 premières CPS sélectionnées selon le pouvoir discriminant sont données par le tableau 1. Les CPS sont numérotées selon l'ordre décroissant des variances (la première CPS est celle de plus grande variance).

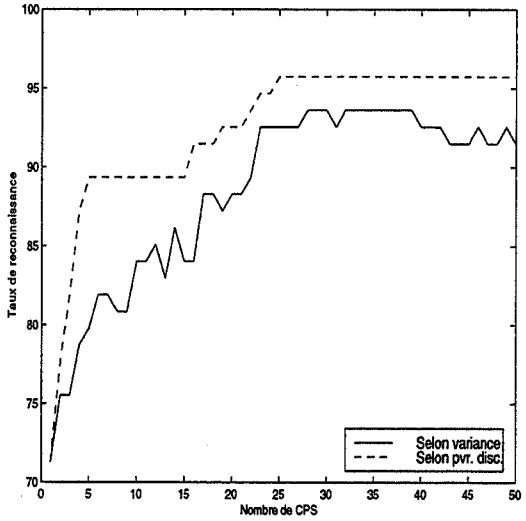


Figure 2: Taux de reconnaissance sur les données de développement, pour les deux méthodes de sélection des CPS.

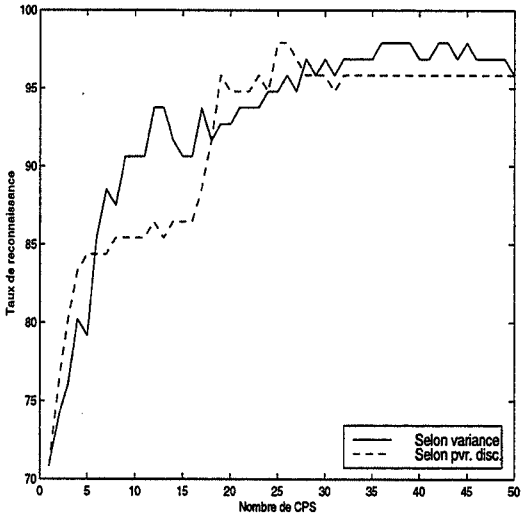


Figure 3: Taux de reconnaissance sur les données de test, pour les deux méthodes de sélection des CPS.

6. INTERPRÉTATION DES RÉSULTATS

6.1. *Comparaison des deux méthodes de sélection*

Avec les deux méthodes de sélection, on observe une stabilisation des performances avec un même nombre de CPS (voir figure 3). Cela est confirmé par le tableau 2 où la liste (réordonnée selon les variances) des 25 premières CPS choisies selon le pouvoir discriminant, correspond approximativement aux CPS de plus grande variance.

Table 1: Liste des 25 premières CPS sélectionnées selon le pouvoir discriminant (présentées selon l'ordre de sélection par le pouvoir discriminant).

NUMEROS DE CPS								
1	6	12	10	32	43	33	16	20
22	26	2	4	7	14	19	15	
5	17	8	9	25	28	11	23	

Table 2: Liste des 25 premières CPS sélectionnées selon le pouvoir discriminant (présentées selon l'ordre décroissant des variances).

NUMEROS DE CPS								
1	2	4	5	6	7	8	9	10
11	12	14	15	16	17	19	20	
22	23	25	26	28	32	33	43	

Cependant la sélection selon le pouvoir discriminant présente de moins bonnes performances entre 6 et 17 CPS. Cela peut s'expliquer par le fait que l'évolution des performances obtenues par cette technique, a une tendance très irrégulière, présentant des intervalles de très faible accroissement des performances. On note comme sur les données de test, un palier pour les performances sur les données de développement, entre 6 et 16 CPS (voir figure 2).

6.2. *Relation entre les CPS et les coefficients cepstraux*

L'analyse cepstrale a été introduite à l'origine, pour la décomposition source/filtre en traitement de la parole. Les coefficients cepstraux sont maintenant utilisés comme paramètres acoustiques en reconnaissance de la parole [d'Al91] [Mam96]. Le cepstre réel sur une échelle linéaire, se définit comme la Transformée en Cosinus du log-module du spectre.

La figure 4 donne une représentation des premières corrélations en valeurs absolues, entre les CPS et les coefficients cepstraux. Ces corrélations sont obtenues à partir des données d'apprentissage. Cette figure met en évidence un très net couplage le long de la diagonale, sauf pour le cadran $[3-7, 3-7]$. Quantitativement, les cases sombres sur la figure correspondent à des coefficients de corrélations en valeurs absolues, compris entre $0,8$ et $0,9$.

Cette similitude entre CPS et coefficients cepstraux sur l'ensemble parole et musique confondues, ne se traduit pas de façon directe du point de vue de la discrimination parole/musique. En effet, pour k inférieur à environ 50, les k premiers coefficients cepstraux fournissent systématiquement de moins bonnes performances que les k CPS de plus grandes variances. Par exemple pour $k = 25$, les coefficients cepstraux ne permettent d'atteindre qu'à peine 90% de segments correctement classifiés, alors que les CPS présentent un taux de reconnaissance d'environ 95%.

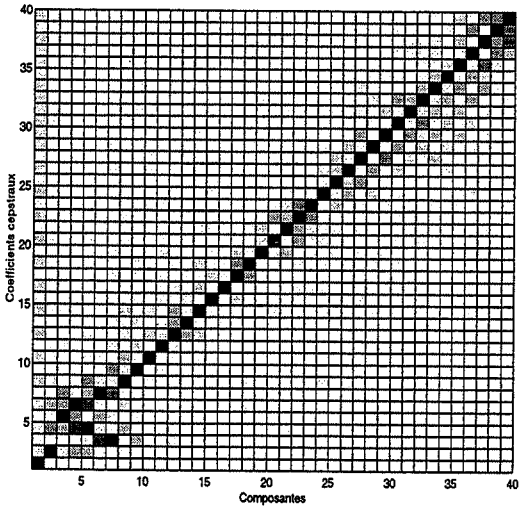


Figure 4: Coefficients de corrélations en valeurs absolues, entre les 40 CPS de plus grandes variances et les 40 premiers coefficients cepstraux.

7. CONCLUSION

Cette étude montre que l'application d'outils simples d'analyse de données et de modélisation statistique, permet d'obtenir des performances satisfaisantes en classification parole/musique d'extraits radiophoniques. Cette approche nous servira de méthode de référence, que nous comparerons à des techniques plus élaborées dans nos futurs travaux.

BIBLIOGRAPHIE

[d'Al91] d'Alessandro C. & Demars C., *Représentations temps-fréquence du signal de parole*, Traitement du signal, Vol. 9, No 2, pp 153-173, Juillet 1991.

[Ked86] Kedem B., *Spectral Analysis and Discrimination by Zero-Crossings*, Proceedings of the IEE, Vol. 74, No. 11, pp 1477-1493, November 1986.

[Mag97] Magrin-Chagnolleau I., *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte*, Thèse de Doctorat de l'ENST, Janvier 1997.

[Mam96] Mammone R. J., Ramachandran R.P., & Zhang X., *Robust Speaker Recognition : A Feature-based Approach*, IEEE Signal Processing Magazine, pp 58-71, September 1996.

[Sap90] Saporta G. *Probabilités Analyse de données et Statistique*, Technip, 1990.

[Sau96] Saunders J. *Real-Time Discrimination of Broadcast Speech/Music*, ICASSP-96, Vol. 2, pp 993-996, November 1996.

[Sch97] Scheirer E. & Slaney M., *Construction and Evaluation of a Multifeature Speech/Music discriminator*, ICASSP-97, Vol. 2, pp 1331-1338, April 1997.

Utilisation de contraintes articulatoires pour l'extraction automatique des formants

Alain Soquet[1] et Véronique Lecuit[2]

[1] Collaborateur scientifique FNRS.
[2] Projet de la Communauté Française de Belgique, ARC 93/98-168.
Laboratoire de Phonétique Expérimentale,
Université Libre de Bruxelles,
av. F.-D. Roosevelt, CP110,
B-1050 Bruxelles– Belgique
Tél.: (32 2) 650 49 12 – Fax: (32 2) 650 20 07 – e-mail asoquet@ulb.ac.be

ABSTRACT

Formants are important cues in the study of speech. Unfortunately, the problem of automatic extraction of these cues is not solved. In this paper we investigate a way to improve formant extraction reliability by using articulatory constraints. Our work is an attempt to tackle this problem in a different way: instead of trying to extract the formant frequencies directly in the acoustic domain, we propose to derive them from the area function of a vocal tract model which is optimized to show similar spectral characteristics as the analyzed signal.

1. INTRODUCTION

Le problème de l'extraction automatique des fréquences des formants est causé à la fois par la difficulté d'estimer les résonances du conduit vocal à chaque instant, l'extraction de formant, et par la difficulté de suivre l'évolution dans le temps des formants, le suivi de formants.

Au niveau de l'extraction de formant, chaque fenêtre d'analyse est traitée isolément sans tenir compte du résultat obtenu dans les fenêtres précédentes ou suivantes. Parmi les méthodes les plus intéressantes, on peut noter les approches du type analyse-synthèse [Oli71], analyse LPC [McC74], analyse cepstrale [Sch70], analyse statistique [Nie85]. Pour le suivi de formants, les méthodes essayent de détecter et corriger les erreurs des systèmes d'extraction en utilisant des règles heuristiques [McC74], des modèles statistiques [Kop86] ou des techniques dérivées du traitement d'image [Lap96].

Notre objectif ici est d'utiliser un nouveau type de contrainte, non plus basé sur des critères acoustiques, tels ceux utilisés par [McC74], [Kop86], et [Lap96], mais basé sur le modèle de production. Le modèle de production peut être utilisé de deux façons dans un système d'extraction de formants: (i) soit pour le suivi, on utilise alors un système d'inversion acoustico-articulatoire pour obtenir des trajectoires articulatoires à partir des formants extraits. L'étude de ces trajectoires permet de juger de la qualité de l'extraction; (ii) soit pour l'extraction, ou l'on dérive les formants à partir de la fonction d'aire d'un modèle du conduit vocal, optimisée pour avoir les mêmes caractéristiques spectrales que le signal analysé. C'est cette approche que nous présentons ici.

2. METHODE

La méthode que nous présentons ici consiste à obtenir, à partir d'indices acoustiques, à l'aide de l'inversion acoustico-articulatoire, une configuration articulatoire de laquelle on extrait les fréquences des formants. Pour chaque fenêtre d'analyse, les paramètres articulatoires sont modifiés de manière à minimiser l'erreur entre les indices acoustiques dérivés du modèle articulatoire et les indices acoustiques calculés à partir du signal (figure 1). Cette opération n'est possible que si l'on peut calculer les fréquences des formants et les indices acoustiques à partir du modèle articulatoire.

Le calcul des fréquences des formants à partir du modèle articulatoire se fait par la méthode variationelle [Jos94]. Les indices acoustiques sont calculés à partir de la fonction de transfert du conduit vocal. La fonction de transfert du conduit vocal est elle-même déduite des fréquences des formants [Fan85].

Dans le cadre de cette étude, le modèle articulatoire utilisé est le modèle à régions distinctives (DRM) [Mra88], les indices acoustiques sont les coefficients cepstraux Mel (MFCC).

La figure 1 illustre la mise en œuvre de la méthode. Dans un premier temps, les MFCC mesurés dans le signal de parole sont comparés aux MFCC contenus dans une table reprenant des couples MFCC-DRM. Les paramètres du DRM minimisant la distance euclidienne entre les MFCC sont utilisés comme paramètres initiaux. Les paramètres sont ensuite ajustés par optimisation de manière à minimiser la distance euclidienne entre les MFCC. Le jacobien de la transformation entre les deux espaces est évalué par perturbation. Après optimisation, les formants calculés à partir du DRM sont considérés comme les formants extraits du signal.

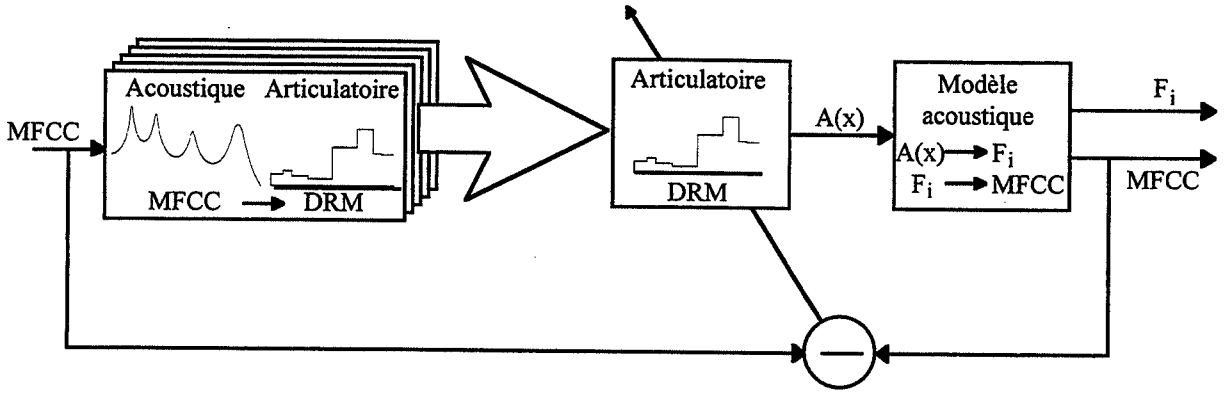


Figure 1: Méthode d'extraction des formants par inversion acoustico-articulatoire.

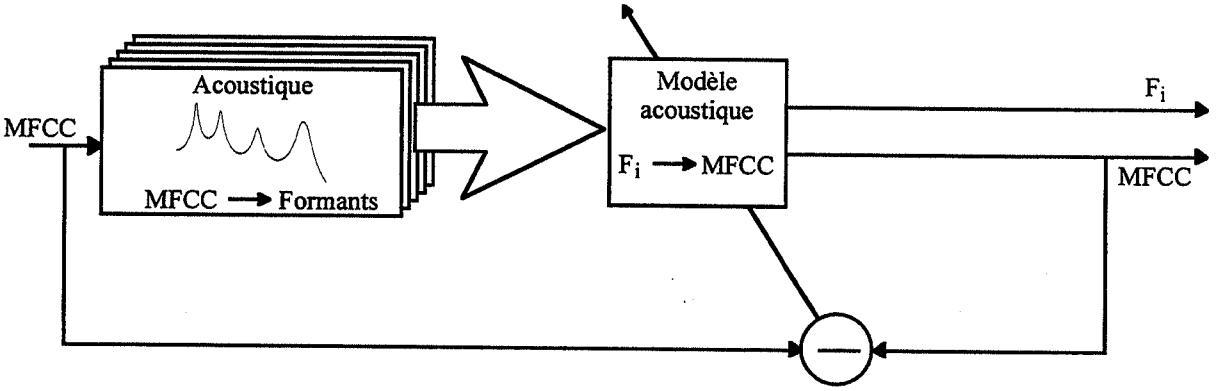


Figure 2: Méthode d'extraction des formants par optimisation des fréquences des formants.

A titre de comparaison, nous avons développé une procédure similaire mais n'utilisant pas de modèle articulatoire. Les paramètres sont dans ce deuxième cas directement les fréquences des formants de manière similaire à [Oli71]. Cette approche est illustrée à la figure 2. Les MFCC sont comparés aux MFCC contenus dans une table comprenant des couples MFCC-formants. Les formants correspondant au meilleur choix participent ensuite au processus d'optimisation. Les formants issus de ce processus d'optimisation sont considérés comme étant les formants extraits du signal.

3 Resultats et discussion

Les deux méthodes présentées ci-dessus, ainsi que la méthode d'extraction par LPC ont été intégrées dans un logiciel. Ce logiciel nous permet de faire une étude comparée des méthodes de manière souple. La référence est obtenue par correction manuelle des formants obtenus par LPC.

Les figures 3 et 4 illustrent le comportement des différentes méthodes LPC (avec correction manuelle), optimisation des paramètres articulatoires par inversion acoustico-articulatoire et optimisation des fréquences des formants, respectivement de haut en bas. La figure 3 présente les résultats pour une partie de phrase extraite de la TIMIT (DAB0 SA1), et la figure 4 pour la séquence de voyelles [aœioy] prononcée par un locuteur masculin.

L'utilisation du modèle de production permet d'avoir un système d'extraction relativement indépendant du nombre de formants que l'on désire extraire. Un des problèmes des méthodes classiques de type LPC est le choix de l'ordre du filtre. Si on surdimensionne le filtre, on augmente le nombre d'erreurs d'insertion. Si on sous-dimensionne le filtre, on favorise les erreurs d'omission. De plus, pour assurer des performances valables, il est nécessaire d'adapter l'ordre du filtre d'une fenêtre d'analyse à l'autre.

Un autre problème des méthodes basées sur l'acoustique provient du fait chaque insertion ou omission implique un suivi de formant incorrect. Or aucune méthode ne permet de détecter de manière fiable les erreurs. Au contraire, le calcul des formants à partir d'une fonction d'aire permet d'identifier les formants de manière fiable.

De plus, si l'on considère le problème des formants proches comme par exemple le F1 et F2 du [u] et F2 et F3 du [y] [Soq95] on s'apperçoit que, par contre, au niveau de l'enveloppe spectrale il n'y a pas de confusion

possible. D'où l'intérêt d'utiliser une méthode d'optimisation basée sur les MFCC qui représentent l'ensemble du spectre.

Les problèmes liés à la deuxième méthode, l'extraction par optimisation des fréquences des formants apparaissent clairement dans les deux exemples des figures 3 et 4. Etant donné qu'il n'y a pas de contrainte explicite sur les formants, le système ajuste simplement ces paramètres pour diminuer sa fonction coût, en l'occurence la distance euclidienne entre les MFCC. On observe donc des problèmes dans les situations critiques classiques, telles que la proximité entre deux formants successifs. Par construction, ce genre de problème ne se présente pas dans la méthode mettant en jeu le modèle de production qui contraint les formants à des valeurs plausibles.

Le problème principal lors de l'utilisation d'un modèle de production dans un système de traitement automatique est lié à l'adaptation au locuteur. Dans notre cas, cette adaptation s'est faite par le choix d'une table adéquate. Cependant, il n'est pas nécessaire de créer des tables adaptées à chaque locuteur, une simple distinction homme-femme nous a permis d'appliquer la méthode de manière satisfaisante.

CONCLUSION

Dans cette étude, nous avons proposé une méthode d'extraction des fréquences des formants qui consiste à obtenir, à partir d'indices acoustiques, à l'aide de l'inversion acoustico-articulatoire, une configuration articulatoire de laquelle on extrait les fréquences des formants. Pour chaque fenêtre d'analyse, les paramètres articulatoires sont modifiés de manière a minimiser l'erreur entre les indices acoustiques dérivés du modèle articulatoire et les indices acoustiques calculés à partir du signal.

BIBLIOGRAPHIE

[Fan85] Fant G. (1985), "The vocal tract in your pocket calculator", STL-QSPR, 2-3, pp. 1-19.

[Jos94] Jospa P. (1994) "Formulation variationelle du lien acoustico-articulatoire", Actes des 20èmes Journées d'Etude sur la Parole, Trégastel, pp. 113-118.

[Kop86] Kopec G.E (1986), "Formant tracking using hidden Markov Models and vector quantization", *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. ASSP-34, 4, pp. 709-729.

[Lap96] Laprie Y. et Berger M.O. (1996) "Cooperation of regularization and speech heuristics to control automatic formant tracking", *Speech Communication*, Vol.19, pp. 255-269

[McC74] McCandless N. (1974) "An algorithm for automatic formant extraction using linear prediction of speech", *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. ASSP-22, pp. 135-141.

[Mra88] Mrayati M., Carré R. et Guérin B. (1988) "Distinctive regions and modes: a new theory of speech production", *Speech Communication*, Vol.7, pp. 257-286

[Nie85] NiederJohn R.J., M. Lahat (1985), "A zero-crossing consitency method for formant tracking of voiced speech in high noise levels". *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. ASSP-33, 2, pp. 349-355.

[OLi71] Olive J.P. (1971), "Automatic formant tracking by a Newton-Raphson technique", J.Acoust. Soc. Am., Vol. 50, 2, pp. 661-670.

[Sch70] Schafer R.W. et L.R. Rabiner (1970), "System for automatic formant analysis of voiced speech", J.Acoust. Soc. Am., Vol. 47, 2, pp. 634-648.

[Soq85] Soquet A. (1985), "A cooperative approach to formant extraction", Proceedings ICPhS-95, Stockolm, Suède, Vol. 4, pp. 448-451.

[Wel98] Welling L. et H. Ney (1998), "Formant estimation for speech recognition", *IEEE Trans. Speech, Audio Processing*, Vol. 6, 1, pp. 36-48.

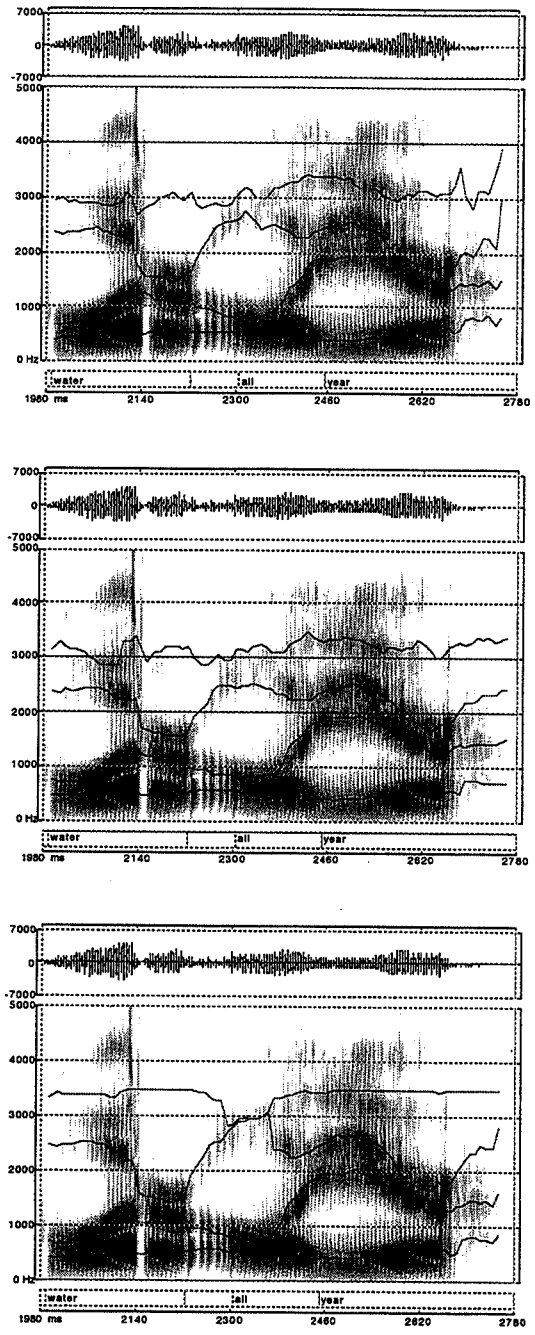


Figure 3: Comportement des différentes méthodes: LPC (avec correction manuelle), optimisation des paramètres articulatoires par inversion acoustico-articulatoire et optimisation des fréquences des formants, respectivement de haut en bas pour une partie de phrase extraite de la TIMIT (DAB0 SA1).

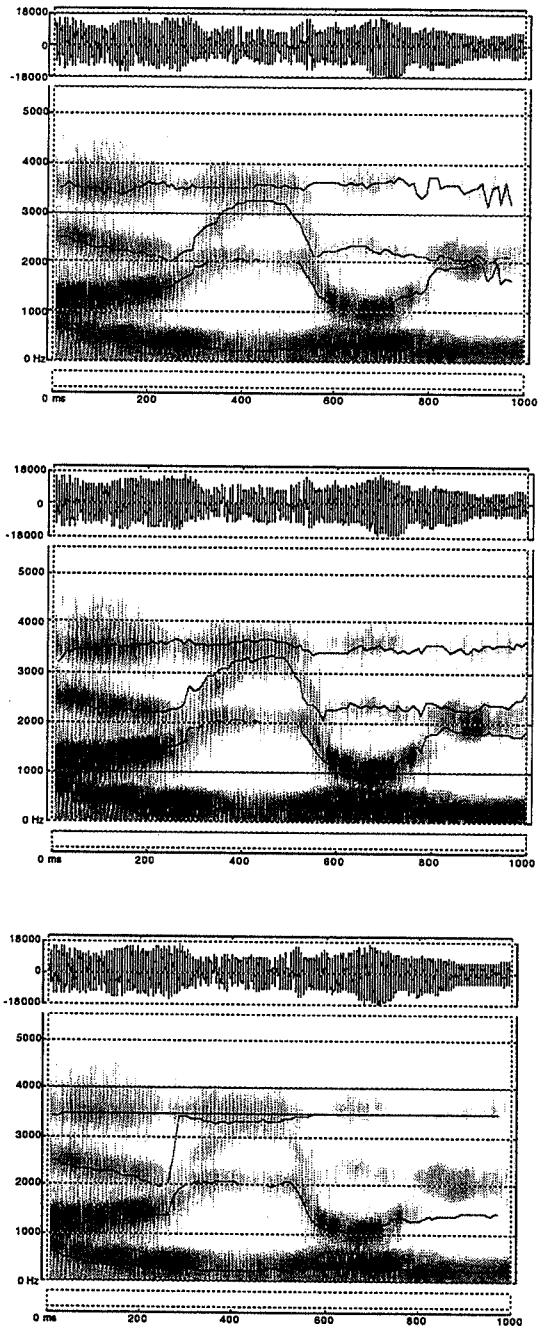


Figure 4: Comportement des différentes méthodes: LPC (avec correction manuelle), optimisation des paramètres articulatoires par inversion acoustico-articulatoire et optimisation des fréquences des formants, respectivement de haut en bas pour la séquence de voyelles [aœioy] prononcée par un locuteur masculin.

Décomposition spectro-temporelle de syllabes et de mots superposés : Résultats préliminaires

Laurent Varin et Frédéric Berthommier

Institut de la Communication Parlée / INPG
46, av. Félix Viallet, 38031 Grenoble cedex 01, FRANCE
{varin,bertho}@icp.grenet.fr

ABSTRACT

We extend the application of a previous modelling technique able to perform stationary double vowel segregation [VB97]. Using the same principle, we decompose non-stationary signals such as mixtures of syllables and words from their time-frequency representations. Preliminary results on double identification performance are shown together with a reduction method of matrix representation.

1. INTRODUCTION

Lorsque plusieurs signaux de parole sont produits simultanément et se superposent, la perception humaine s'adapte et l'intelligibilité de ces signaux reste satisfaisante (il s'agit de l'effet 'cocktail party'). Les études psychoacoustiques ont révelé l'existence d'un phénomène de ségrégation dit primitif [Bre90] qui conduit à la perception de 'flux auditifs' clairement séparés, ceci à partir de mélanges de signaux complexes qui paraissent inextricables si on examine leur spectrogramme. Cette capacité s'appuie sur des indices tels que l'harmonicité, la simultaneité de début ou de fin des composantes, ou bien encore la localisation spatiale. Les modèles de ce processus de ségrégation de sources sonores font appel à des modes de traitement et de représentation des signaux qui visent à extraire, puis à regrouper les composantes du signal émises par une source physique donnée (réunis sous l'appellation CASA, 'Computational Auditory Scene Analysis' ; certains développements sont réalisés dans le cadre du contrat Européen TMR SPHEAR). Cette fonction s'appliquerait sur tous les signaux complexes, y compris la parole, et nous pouvons donc nous attendre à ce qu'elle sous-tende l'effet 'cocktail party'.

Sans remettre en question la validité de cette approche, nous pensons que les processus de ségrégation sont susceptibles d'impliquer directement le niveau d'identification de la parole (que nous appelons aussi niveau phonétique). Visualisée dans le plan temps-fréquence, la structure des signaux de parole est très spécifique : c'est une succession de bruits consonantiques et de noyaux vocaliques où les formants apparaissent très nettement, portés par les composantes harmoniques. Ces distributions énergétiques structurées sont encodées dans le signal par le tractus vocal, par résonance du conduit et lors des mouvements articulatoires. Dans un système de reconnaissance automatique, lors de la phase de décodage, les formes présentes dans le plan temps-fréquence sont appariées aux formes prototypiques connues par le système, accumulés au cours de la phase d'apprentissage. Le mode de représentation varie suivant les systèmes, et les modèles probabilistes, fondés sur les HMM, organisent et utilisent les données à partir d'une description statistique Gaussienne. Ils codent chaque classe à l'aide d'un vecteur moyen associé à une matrice de variance covariance. Avec un HMM 'standard', que nous prendrons comme référence des processus de reconnaissance automatique, les états spectraux stationnaires sont étiquetés, et les dépendances temporelles sont codées par les dérivées temporelles et par les probabilités de transition. Ces informations sont supposées nécessaires et suffisantes pour réaliser le décodage acoustico-phonétique, ce qui revient à tenir compte des informations encodées par le mécanisme de production en s'appuyant uniquement sur l'image acoustique résultante.

Cette représentation a priori, mémorisée, des signaux de parole est susceptible d'être utilisée au cours du processus de perception humain, soit pour réaliser la ségrégation complète de signaux (Hypothèse forte de la ségrégation fondée sur les schémas), soit au sein d'un système multi-niveaux (Hypothèse du couplage). L'hypothèse forte s'appuie sur le point de vue développé par Remez et coll. [Rem+94], selon lequel les niveaux intermédiaires susceptibles d'être impliqués dans l'ASA ne joueraient aucun rôle pour la parole. Une argumentation opposée a été développée par Barker et Cooke [BC97], qui s'appuient sur l'utilisation des mêmes types de signaux. Ceux-ci sont dénués de la presque totalité de leurs indices primitifs (SWS, Sine Wave Speech, qui restitue les trajectoires formantiques et élimine la contribution de la source glottique, en particulier l'harmonicité). Ces signaux sont identifiables par les auditeurs humains, et de surcroît peuvent être reconnus par un HMM 'standard'. Mais dans la condition de superposition des SWS, l'intelligibilité de

la source dite 'non dominante' diminue fortement (par rapport à des signaux normaux). Nous défendrons l'hypothèse du couplage, car nous pensons que les informations primitives et phonétiques sont analysées dans des niveaux différents et que toutes deux contribuent au processus de ségrégation. Nous montrerons quelques éléments en faveur de ce compromis.

Afin d'instancier le niveau de ségrégation phonétique, nous développons un modèle basé sur des représentations statistiques et sur l'application d'un algorithme de décomposition fondé sur la maximisation de la vraisemblance des signaux résultants. Cet algorithme, dont le principe a été proposé par Varga et Moore [VM90], a été appliqué dans les cas où la parole interfère avec un bruit qui peut être modélisé par un faible nombre d'états HMM. Cette méthode présente l'inconvénient d'être combinatoire, et comme amélioration, nous articulons le processus de décomposition autour de l'effet de dominance. A un moment donné, l'élément dominant est reconnu par une méthode robuste (et faisant éventuellement appel aux indices primitifs). La décomposition est amorcée en re-traitant le signal à partir de la connaissance dont on dispose sur l'élément dominant. A ce stade, d'autres algorithmes réalisent une soustraction spectrale afin d'identifier le résidu. Mais le processus de décomposition réalise une ségrégation à proprement parler en partageant le signal d'entrée en deux signaux identifiables.

Nous mettrons en évidence des points d'entrée venant du niveau primitif, notamment ceux porteurs d'informations sur l'organisation temporelle et sur les niveaux d'intensité relatifs.

2. LA SEGREGATION DE DOUBLE VOYELLES

2.1 *L'effet de dominance*

Dans un premier temps, nous ne considérons que des superpositions de voyelles stationnaires à moyen terme (50 ms). Ce paradigme de reconnaissance des double voyelles a été utilisé dans le cadre d'une comparaison expérimentation/modélisation, de façon à contrôler l'introduction des indices primitifs (ici l'harmonicité et la fréquence fondamentale) et à évaluer leur rôle dans le processus de ségrégation. Ces stimuli possèdent les deux types d'attributs (primitifs et phonétiques), et il est difficile de conclure quant au niveau de ségrégation effectivement mis en jeu. En marge de nombreuses observations faites à partir de ce paradigme, il a été constaté que lorsque des paires de voyelles sont présentées à des sujets, l'une d'elle est dominante et immédiatement identifiable. Nous sommes donc dans une situation de type figure sur fond, dans laquelle la figure, saillante perceptivement, est identifiée avec succès alors que le fond, masqué, est responsable du taux d'erreur de double identification.

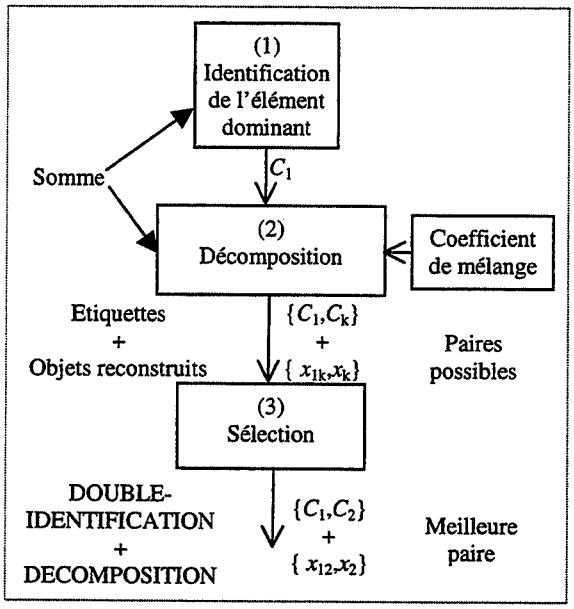


Figure 1 : Les étapes du modèle de ségrégation

Le phénomène de dominance semble du à plusieurs facteurs. Nous pouvons incriminer les deux types d'attributs, primitifs et phonétiques, ainsi qu'un phénomène attentionnel. Mais, pour l'élaboration de notre modèle, nous considérons qu'il est lié à l'identification des mélanges. Lorsqu'un classifieur est entraîné à l'aide d'objets isolés, et que la reconnaissance s'effectue sur un mélange de deux ou plusieurs objets, ce mélange reçoit généralement l'étiquette de l'élément dominant. Bien entendu, le rôle des attributs de bas niveau n'est pas éliminé, et en particulier, le niveau d'intensité relatif est un paramètre essentiel du paradigme des double voyelles, qui permet de contrôler la saillance perceptive de la voyelle dominante.

2.2 *L'étape de décomposition*

La formule de décomposition donne la solution d'un système d'équations à n inconnues qui sont les fractions d'énergie transmises par chaque canal (un canal = une dimension de l'espace d'entrée), pour chacun des éléments en présence. Cette fraction est pondérée par une variable globale qui correspond au niveau d'intensité relatif des deux éléments en présence, ou coefficient de mélange. Celui-ci peut être : (1) fixé a priori (2) évalué par une méthode dite centrale (C-eval), par comparaisons avec les moyennes des prototypes stockés, (3) évalué à partir d'un mécanisme primitif (P-eval). Cette dernière méthode est compatible avec la notion de couplage entre niveaux, le rôle des niveaux intermédiaires étant d'attribuer une sonie à chacun des flux auditifs.

2.3 *Le modèle de ségrégation*

Le modèle comporte trois étapes (Figure 1) :

(1) Identification robuste, donnant l'étiquette C_1 de l'élément dominant, par analyse discriminante linéaire.

(2) La décomposition en paires possibles $\{C_1,C_k\}$ (avec k≠1). La méthode du maximum de vraisemblance est appliquée pour reconstruire des spectres x_1 et x_2 par une maximisation du produit $P(x_1|C_1)*P(x_2|C_2)$ tel que $x = \alpha.x_1+(1-\alpha).x_2$, à partir du spectre d'entrée x, voir [VB97] pour le détail des calculs.

(3) Sélection de la meilleure paire reconstruite, donnant les deux étiquettes $\{C_1,C_2\}$ ainsi que les spectres les plus probables.

2.4 Simulations

Nous avons effectué des tests sur une base de voyelles synthétiques du français (/aeiouy/), dont la fréquence fondamentale varie entre 100 et 200 Hz. Le signal est représenté par un spectre sur 100-4700 Hz avec 32 canaux selon une échelle Bark. Les erreurs produites sont soit des distorsions spectrales, soit des erreurs d'étiquetage. Elles ont trois origines : (1) l'évaluation du coefficient de mélange est une méthode heuristique (2) l'algorithme de décomposition produit les spectres les plus probables, et non pas les spectres d'entrée (3) les signaux sont supposés être spectralement additifs, ce qui n'est pas le cas (à cause des incohérences de phase). Malgré ces imperfections, cet algorithme est remarquablement efficace lorsqu'il est appliqué à la ségrégation des double voyelles (Table 1). La ségrégation à partir du coefficient de mélange évalué par le niveau primitif (P-eval) reste robuste lorsque la différence d'intensité entre les deux sources est importante. Les bons scores obtenus avec un coefficient constant montrent aussi que la décomposition est robuste aux erreurs d'évaluation de ce coefficient, lorsqu'il est fixé à une valeur moyenne.

Table 1 : Taux d'identification des paires de voyelles, d'après leur niveau d'intensité relatif exprimé en dB RMS. Le coefficient constant correspond à une valeur moyenne de 0dB RMS. Les indices primitifs utilisés pour l'évaluation P-eval du coefficient de mélange sont le F0 (fréquence fondamentale) et l'ITD (différence interaurale de temps).

Coefficient de mélange α	Niveau relatif			
	0dB	6dB	12dB	18dB
Constant (α=0.65)	98%	88%	60%	22%
C-eval	96%	92%	73%	31%
P-F0-eval	95%	89%	75%	26%
P-ITD-eval	97%	92%	75%	53%

3. SEPARATION DE SIGNAUX NON STATIONNAIRES

3.1 La ségrégation de mots alignés

Nous avons effectué un test de séparation de mots alignés en utilisant le même principe. Dans la base de données 'Numbers 95', nous avons extrait 1000 exemples (en moyenne) de chacun des cinq mots « one », « two », « three », « four », « five » afin de constituer la base d'apprentissage. La base de test comprend 11 exemples supplémentaires de chaque mot. Une image temps-fréquence de chaque mot est obtenue en calculant un spectre Mel à 20 canaux, sur des fenêtres successives de 30ms, et la taille de cette image est fixée à 20*10. Le grand vecteur d'entrée temps-fréquence et les matrices de variance-covariance correspondantes sont traitées de la même manière que précédemment (chaque carreau temps-fréquence étant associé à une dimension de l'espace d'entrée).

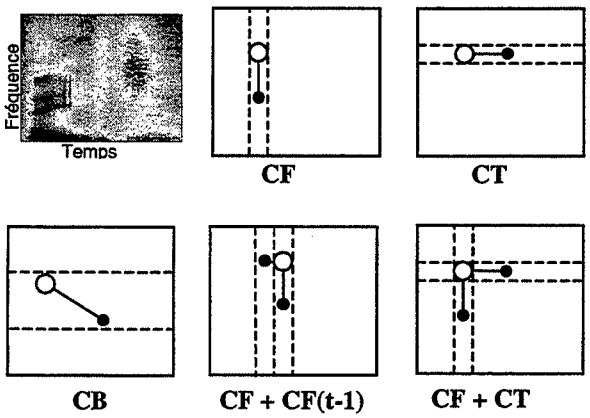


Figure 2 : Corrélations dans une image temps/fréquence du signal. Exemples : corrélation entre deux valeurs (o) et (•). **CF** : de fréquences différentes. **CT** : de fenêtres de temps différentes. **CB** : dans une bande de fréquences. **CF+CF(t-1)** : de fréquences différentes ou de la fenêtre de temps précédente. **CF+CT** : de fréquences différentes ou de fenêtres de temps différentes.

A l'aide de l'analyse discriminante linéaire, les mots isolés sont identifiés à 88% dans la base d'apprentissage et 89% dans la base de test. Les paires de mots superposés à 0dB RMS sont identifiées à 81%. Pour 95% des paires, l'un des mots est identifié après la première étape d'identification du dominant. Un élément au moins est identifié dans 99% des cas, après la sélection de la meilleure paire reconstruite. Ces résultats sont bons, mais la représentation que nous utilisons n'est pas appropriée, car l'estimation des matrices de variance covariance nécessite un très grand nombre d'exemples. En effet, ces matrices représentent les corrélations entre toutes les paires de carreaux dans le plan temps-fréquence. Certaines composantes sont manifestement inappropriées, par exemple celles exprimant la covariance entre deux points à la fois éloignés spectralement et temporellement. De plus, même si une covariance non nulle est observée pour de telles paires, elle peut être le résultat d'une corrélation indirecte, et par conséquent redondante. Bien entendu, cette représentation pose d'autres problèmes liés à l'alignement des signaux (ici donné a priori) et à leur renormalisation.

3.2 Creuser la matrice de variance-covariance

Par conséquent, la stratégie que nous proposons consiste à étudier plus précisément les modes de réduction de la grande matrice temps-fréquence dont nous proposons une systématisation (Figure 2) :

CF : Corrélations entre des carreaux de fréquences différentes dans une même fenêtre temporelle.

CT : Corrélations entre des carreaux éloignés dans le temps, mais de fréquence identique (auto-corrélation).

Nous creusons la grande matrice en mettant arbitrairement à 0 les termes non pris en compte. Afin de préserver une information sur l'évolution temporelle, nous ne conservons que les termes CF + CT. D'autres combinaisons existent pour incorporer l'information fréquentielle et temporelle tout en creusant la matrice.

3.3 Ségrégation de double syllabes

Nous avons testé ce modèle sur une base monolocuteur de 6 syllabes VC du Français (/ad/, /ag/, /id/, /ig/, /ud/, /ug/), avec seulement 14 exemples par classe. Contrairement au cas précédent, le faible nombre d'exemples ne permet pas une estimation fiable des matrices de covariance. Il est à priori préférable de n'utiliser qu'une partie de l'information. Ainsi, nous constatons que les différents modes de réduction donnent des résultats satisfaisants et le meilleur score est obtenu par CF+CT, avec 58%, contre 47% avec les matrices pleines (Table 2).

Table 2 : Taux d'identification (%) des paires de VC. L'identification du dominant fait appel à l'analyse discriminante linéaire. Nous indiquons le taux de double identification par la méthode 2D-max, qui donne directement le second candidat, puis les taux obtenus avec la matrice pleine et différents types de matrices creuses.

Méthode	Reco dominant	Double reco
(2D-max)	88%	18%
Matrice pleine	88%	47%
Matrice diagonale	86%	43%
CT	87%	44%
CF	86%	53%
CT+CF	87%	58%

4. CONCLUSION

Il s'agit de développer un modèle de décomposition temps-fréquence de signaux de parole superposés. Pour cela, nous devons intégrer notre modèle dans un système de reconnaissance de la parole continue. De façon à guider un système de décomposition et de reconnaissance de type HMM, nous proposons que ces informations d'alignement (liées à l'organisation temporelle des flux) soient extraites par le niveau primitif. Le problème de la détermination de l'alignement est l'équivalent temporel de l'évaluation des niveaux d'intensité relatifs des signaux en présence, que nous avons traité dans le domaine spectral pour la ségrégation des double voyelles.

Avec un modèle de type HMM, un signal non stationnaire est traité comme une succession de segments stationnaires auquels sont affectés des probabilités d'appartenance à une classe, et qui sont enchaînés à l'aide de l'algorithme de Viterbi, intégrant aussi les probabilités de transition. Par conséquent, un HMM ne fait pas appel à une description explicite des corrélations spectro-temporelles par paires, en dehors d'un nombre très limité de fenêtres temporelles (les dérivées peuvent néanmoins être considérées comme un petit sous-ensemble de cette description complète).

Des améliorations de ce modèle classique sont recherchées en explicitant l'ensemble des corrélations nécessaires et suffisantes, et surtout en identifiant le niveau dans lesquelles elles sont extraites et prises en compte. En effet, les évolutions et les corrélations temporelles d'enveloppes RMS (ici par canal spectral) sont des indices primitifs qui ne sont pas spécifiques des signaux de parole, bien que pouvant sous-tendre le processus de reconnaissance des consonnes (voir [Apo+97]). Cette préoccupation est au centre d'une démarche récente qui consiste à rechercher une telle amélioration en réduisant la redondance de la représentation HMM classique (déjà très réduite par rapport à notre grande matrice). Une réduction est réalisée avec succès dans le modèle de reconnaissance en sous-bandes [BD96], qui ne tient compte que des corrélations spectrales au sein d'un nombre limité de domaines (3 ou 4) qui codent chacun un attribut formantique en moyenne. D'autre part, les dépendances temporelles peuvent être introduites d'une façon contrôlée via une étape de pré-traitement par extraction et lissage des enveloppes RMS au niveau syllabique.

5. BIBLIOGRAPHIE

[Apo+97] Apoux, F. et coll. (1998), ce proc.

[BC97] Barker, J. & Cooke, M. (1997), "Is the sine-wave cocktail party worth attending?", CASA'97, 1-10.

[BD96] Bourlard, B. & Dupont, S. (1996), "A new ASR approach based on independent processing and recombination of partial frequency bands", in proc. of Int. Conf. on Spoken Lang. Proc., Philadelphia, 422-425.

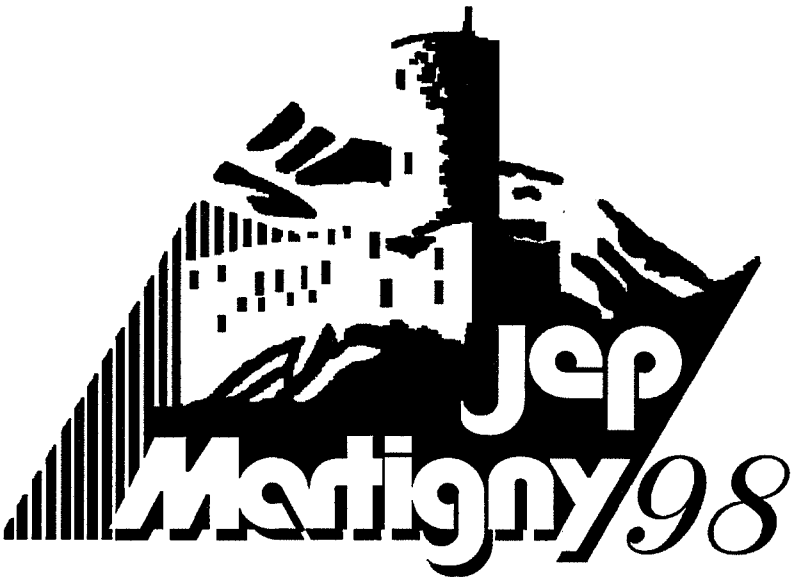
[Bre90] Bregman, A.S. (1990), "Auditory scene analysis", MIT Press.

[Rem+94] Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S. & Lang, J.M. (1994), "On the perceptual organization of speech", Psychological review, 101(1):129-156.

[VB97] Varin, L. & Berthommier, F. (1997), "A probabilistic model of double-vowel segregation", EuroSpeech'97, Rhodes.

[VM90] Varga, A.P. & Moore, R.K. (1990), "Hidden Markov model decomposition of speech and noise", ICASSP-90, 845-848.

Modèles de langage



Un modèle de langage mixte basé sur la similarité des mots dans un système de reconnaissance de parole

Christel Beaujard et Michèle Jardino

LIMSI-CNRS
Université Paris-Sud
BP 133,91403 Orsay Cedex, France
{beaujard,jardino}@limsi.fr

Abstract

Stochasctic n-gram models are commonly used in speech recognition systems. But when the data size is small, more robust models are required. One method, to compensate the sparseness of data, is to gather words into classes. Here we use a local criterion of classification, which takes the similarity of word contexts into account, defined as the relationships between adjacent words. The main advantage of this method is that no class is ambiguous. At the end of the classification, classed and unclassed words can be found. The mixed language model, built from this classification, has been trained from transcribed speech, and then used in the LIMSI speech recognizer to be evaluated in term of recognition rates. A study shows that, even if the mixed language model doesn't give the best recognition rates, the system recognizes better some semantic categories than with a simple word-based model.

1 Introduction

Les modèles de langage probabilistes de type n-grammes de mots sont couramment utilisés dans les systèmes de reconnaissance de la parole. Ces modèles sont bien entraînés lorsque les données d'apprentissage sont conséquentes. Mais lorsque la taille de ces données est réduite comme dans les tâches de dialogue, cas étudié ici, il est nécessaire de créer des modèles de langage plus robustes. Une méthode, pour combler ce manque d'information, est de regrouper des mots dans des classes et de construire des modèles de langage à partir de ces classes.

Il a été montré que des méthodes de classement automatique peuvent être efficaces pour regrouper les mots [Jar96] et [Jar97]. Ces méthodes sont basées sur la notion de minimum de perplexité. Ici, nous proposons un critère de classement qui prend en compte la similarité de contexte des mots. L'avantage principal de cette méthode est la cohérence des mots d'une même classe.

L'approche a déjà été décrite dans [Bea97]. C'est pourquoi nous ne présentons que les grandes lignes du modèle de langage et nous nous focalisons sur les résultats apportés, à partir de données plus conséquentes. Le modèle de langage a été évalué en termes de perplexité et de taux de reconnaissance. Mais les résultats qui sont présentés ici vont au delà des taux de reconnaissance habituels, car nous étudions le taux de mots reconnus dans des catégories sémantiques fixées.

2 Classification automatique

Dans cette section, nous décrivons l'algorithme qui regroupe des mots non étiquetés dans des classes en considérant la similarité des contextes de ces mots.

2.1 *Contextes de mots*

La similarité des mots a déjà été utilisée dans la construction de modèles de langage [Far96], mais cette notion diffère suivant la définition que l'on donne au contexte des mots [Chu90], [Dag95], [Huc95] et [Far96]. Notre approche a pour but d'améliorer les performances du système de reconnaissance du LIMSI. On s'intéresse donc aux interactions entre mots adjacents.

Pour un vocabulaire, $V_L = \{v_1, \ldots, v_L\}$ de taille L, le contexte d'un mot du vocabulaire est l'ensemble des mots rencontrés immédiatement avant et après ce mot dans le texte d'apprentissage.

Plus présicément, on considère le contexte de chaque mot v_i comme l'ensemble des bigrammes gauches et droits qui lui sont associés dans le texte d'apprentissage. On note $\{v_k, v_i\}_k$ l'ensemble des bigrammes gauches de v_i et $\{v_i, v_k\}_k$, l'ensemble de ses bigrammes droits.

2.2 *Critère de classification*

Pour comparer deux mots du vocabulaire v_i et v_j, on définit les ensembles de mots v_k, contextes de v_i et v_j et dont les bigrammes apparaissent plus d'une fois dans l'apprentissage. $CGC(v_i, v_j)$ est l'ensemble des contextes gauches en commun de v_i et v_j et $CDC(v_i, v_j)$ l'ensemble de leurs contextes droits en commun.

Les mots v_i et v_j sont similaires et on leur attribue une même classe, si chacune des quatre sommes de probabilités conditionnelles sur les ensembles de contextes en commun est supérieure à un seuil β, fixé avant le classement.

2.3 *Algorithme de classification*

Les classes sont construites itérativement si le critère de similarité de deux mots est rempli. Au début de la classification, il n'existe encore aucune classe. Tous les mots sont comparés deux par deux. L'ordre de comparaison n'influe pas sur le classement, les mots sont donc pris, ici, dans l'ordre alphabétique. Si deux mots sont considérés comme similaires, ils sont regroupés dans une même classe en fonction de l'état dans lequel ils se trouvent:

- Si les deux mots n'appartiennent encore à aucune classe, une nouvelle classe est créée.
- Si seulement un des deux mots est déjà classé, le deuxième est rajouté dans la classe.
- Si les deux mots sont déjà classés dans des classes différentes, on fusionne les deux classes.

Le nombre final de classes n'est pas connu *a priori*. Il dépend des résultats de chaque itération, du vocabulaire, du texte d'apprentissage et du critère de similarité.

A la fin du classement, il y a des mots classés et d'autres qui ne le sont pas. On observe donc quatre types de bigrammes qui donnent lieu à quatre types d'interaction.

- L'interaction "mot-mot" provient de la succession de deux mots qui n'ont pas été classés.
- L'interaction "classe-mot" correspond à la succession d'un mot classé par un mot qui ne l'est pas.
- Si un mot classé suit un mot non classé, l'interaction qui en résulte est "mot-classe".
- Si les deux mots consécutifs sont classés, l'interaction est appelée "classe-classe".

Le modèle de langage construit à partir de ce classement doit donc tenir compte de ces quatre types d'interaction.

3 Modèle de langage mixte

Le but d'un modèle de langage est de prédire tous les événements possibles connaissant le passé. Même si la taille, L, du vocabulaire est petite, un texte d'apprentissage provenant d'une tâche de dialogue ne peut pas fournir toutes les interactions. Il faut donc interpoler les probabilités des bigrammes non vus.

Pour déterminer ces probabilités d'évéments non observés, on utilise une interpolation non linéaire [Ney94]. Une valeur fixée est déduite de l'occurrence de tous les bigrammes observés et la masse correspondante est redistribuée sur tous les bigrammes possibles.

Les quatre types d'interactions (section 2.3) impliquent quatre types de probabilités conditionnelles de bigrammes basés sur des mots et des classes. Pour incorporer le modèle de langage dans notre système de reconnaissance, il faut qu'il donne les probabilités conditionnelles des bigrammes de mots, c'est à dire les bigrammes basés uniquement sur les mots du vocabulaire. On éclate donc chaque classe en combinant, par un produit, l'information des probabilités conditionnelles de bigrammes mixtes et le poids des mots dans leur classe.

Le modèle de langage est donc construit sur la base du classement présenté dans la section 2 et tient compte de l'état des mots en fin de classement.

4 Description des données

Les données utilisées pour l'apprentissage, le développement et le test proviennent du corpus MASK (Multimodal-Multimédia Automated Service Kiosk) [Lam95]. Le kiosque informe les voyageurs ferroviaires des prestations, des horaires, tarifs, ... des trains en France. Des sujets ont été enregistrés au LIMSI, puis toutes les données ont été transcrites.

Des transcriptions de ces données, 46810 phrases de 9 mots en moyenne, ont été utilisées pour l'apprentissage. Le vocabulaire, constitué des mots du texte d'apprentissage et complété des noms de ville, a une taille de 1531 mots (comprenant les symboles de début et de fin de phrase et le mot inconnu regroupant tous les mots de l'apprentissage non compris dans le vocabulaire (0,2%)). Le corpus d'apprentissage compte 14505 bigrammes de mots différents, soit 0,6% des bigrammes à représenter.

Pour le développement et l'évaluation des modèles de langage, nous avons utilisé les transcriptions et les signaux de paroles des données non utilisées à l'apprentissage. Le développement contient 1864 phrases de 9 mots en moyenne et le test 1789 phrases de 10 mots en moyenne. Dans les deux cas, les mots hors vocabulaire représentent moins de 1%.

5 Résultats

Le modèle de langage mixte a été utilisé par le système de reconnaissance du LIMSI à la place du modèle de langage basé sur les mots, de manière à l'évaluer dans la tâche qu'il doit remplir. Les conditions expérimentales sont identiques pour chacun des modèles, le seul paramètre variant dans le système de reconnaissance est le modèle de langage.

5.1 *Détermination du paramètre β*

β, le seul paramètre du classement, est fixé avant la classification automatique des mots. Il représente le seuil de formation des classes. Pour déterminer sa valeur, on a évalué dans un premier temps les taux de reconnaissance et les perplexités obtenus sur le texte de développement. Ensuite, nous avons comparé finement les résultats de reconnaissance de catégories sémantiques attachées à la tâche. Mais avant de s'intéresser aux résultats de reconnaissance, une brève section décrit les modèles de langage mixtes par leur caractéristiques, en fonction du paramètre de classification.

Caractéristiques des modèles de langage mixtes

Le paramètre β est compris entre 0 et 1. Sa valeur détermine la possibilité de créer ou de compléter des classes. Plus β est petit, plus les classes se forment facilement, mais elles ont moins de cohérence, dans le cas extrême où β est nul, tous les mots du vocabulaire observés à l'apprentissage se retrouvent dans une classe unique. A l'opposé, plus β est grand, plus il est difficile de regrouper les mots. Pour une valeur inférieure à 0,7, la classification automatique regroupe beaucoup de mots dans peu de classes. Et pour β plus grand que 0,9, le critère de classification est trop strict, peu de mots se trouvent classés. La détermination de β se fait donc entre 0,7 et 0,9. Dans cet intervalle, seulement trois valeurs de β sont reportées ici, des valeurs intermédiaires ayant été testées et n'apportant pas de nouvelles informations.

Pour les trois modèles mixtes correspondant aux valeurs de β de 0,7, 0,8 et 0,9, les résultats de classification sont résumés dans le tableau 1. Les taux de mots classés sont calculés sur le vocabulaire et sur la totalité du texte. Le nombre de bigrammes est donné avant et après éclatement des classes (section 3).

	Modèles Mixtes		
	β=0,7	β=0,8	β=0,9
Nombre de classes	46	31	9
Mots classés (% en voc.)	18	11	2
Mots classés (% en tot.)	49	30	8
Nbre de bigrammes	8994	10749	13913
Nbre de bigrammes de mots	87906	47820	16229

TAB. 1 – *Caractéristiques des classifications*

Plus β est petit, plus les classes sont conséquentes et plus le nombre de bigrammes de mots augmente, donc moins il y a de bigrammes de mots non observés à l'apprentissage.

Taux de reconnaissance et perplexités

Les taux de reconnaissance sont calculés par une méthode dynamique d'alignement des mots [Pal90]. Le taux de mots repésente la proportion des mots reconnus sur la totalité du texte, (word accuracy) par rapport à la référence. La perplexité est calculée en considérant les bigrammes de mots donnés par le modèle de langage mixte éclaté.

Le tableau 2 donne les taux de mots reconnus et les perplexités obtenues avec les modèles de langage mixtes pour β variant de 0,7 à 0,9. Le modèle de langage appris directement à partir des bigrammes de mots du corpus d'apprentissage sert de référence, nous l'appellerons modèle de référence.

Dans un premier temps, on remarque que l'évolution de la perplexité suit celle des taux de reconnaissance. Le modèle mixte avec un paramètre égal à 0,7 donne les plus mauvaises performances à la fois en terme de perplexité et en taux de reconnaissance. Le modèle se rapprochant

	Modèle de réf.	Modèles Mixtes		
		β=0,7	β=0,8	β=0,9
Perplexités	**14**	18	15	**14**
Taux de mots(%)	**91,7**	89,6	91,2	91,6

TAB. 2 – *Paramètre β, reconnaissance et perplexités*

le plus du modèle de référence est celui dont le critère de classification est le plus strict (β=0,9).

Aucun modèle n'améliore les résultats de la reconnaissance obtenus avec le modèle de référence. Mais les erreurs, commises par le système, sur des mots ayant peu d'importance pour la tâche de dialogue comme les hésitations, les formules de politesse, etc, comptent autant que celles sur des mots cibles tels que les villes, les jours, les heures, etc. Il est donc intéressant de regarder si les modèles de langage influent de la même manière sur des catégories sémantiques différentes.

Etude de reconnaissance de catégories sémantiques

Pour étudier la reconnaissance de catégories sémantiques, on se fixe des ensembles de mots clés pour la tâche. Les principales catégories sémantiques de la tâche de dialogue sont: les villes et gares, les jours, les mois, les chiffres. Le tableau 3 donne la composition des classes de villes, de mois et de chiffres. La catégorie des jours n'est pas représentée, car aucun des trois classements ne regroupe ces mots. Dans ce tableau, M désigne le nombre de mots classés de la catégorie, C, le nombre de classes contenant ces mots et FA le nombre de mots hors de la catégorie dans ces classes (fausses alertes).

	β=0,7			β=0,8			β=0,9		
	M	C	FA	M	C	FA	M	C	FA
Villes	97	4	19	72	4	3	12	2	1
Mois	12	1	0	11	1	0	2	1	0
Chiffres	18	3	0	11	2	0	0	0	0

TAB. 3 – *Composition des classes de villes, mois et chiffres*

Pour évaluer les modèles sur chaque catégorie, on calcule le taux de reconnaissance de chacun de ses mots pour en déduire le pourcentage moyen de reconnaissance de la catégorie sémantique. Ces taux sont calculés à partir des omissions et des substitutions des mots de la catégorie. Dans le tableau 4 les pourcentages de reconnaissance des cinq catégories sémantiques sont donnés pour chacun des modèles.

Catégories sémantiques	Modèle de réf.	Modèles Mixtes		
		β=0,7	β=0,8	β=0,9
Taux de villes	93,1	89,8	**93,5**	92,6
Taux de jours	85,4	**88,3**	85,4	85,4
Taux de mois	96,1	95,7	**96,3**	96,1
Taux de chiffres	**92,5**	90,6	91,9	**92,5**

TAB. 4 – *Reconnaissance des catégories sémantiques*

Dans ce tableau, on voit que les taux de reconnaissance des catégories sémantiques ne suivent pas l'évolution des taux globaux donnés par le tableau 2. Le modèle mixte avec β=0,8, améliore la reconnaissance des villes et des mois, alors que les jours sont mieux reconnus avec le modèle de langage mixte avec β=0,7. Pour β=0,9, les résultats sont pratiquement équivalents à ceux obtenus avec le modèle de référence exceptés pour les villes pour lesquels le modèle de référence est préférable.

Deux des modèles mixtes ont donc un intérêt, ceux avec β=0,7 et β=0,8. Il ont tous les deux au moins une catégorie pour laquelle ils améliorent la reconnaissance par rapport au modèle de référence. Mais pour l'évaluation, on garde celui dont les taux sont meilleurs pour deux catégories sur les quatre, c'est à dire le modèle de langage mixte dont le paramètre de classification β est égal à 0,8.

5.2 *Evaluation du modèle mixte*

Le modèle de langage mixte, avec β égal à 0,8 est évalué sur le test, décrit dans la section 4. On compare le taux de reconnaissance global et les taux de mots reconnus dans les catégories sémantiques fixées dans la section 5.1.

	Modèle de réf.	Modèle Mixte β=0,8
Perplexités	**14**	15
Taux de mots(%)	**90,4**	90,5
Taux de villes(%)	77,2	**77,9**
Taux de jours(%)	87,6	**93,8**
Taux de mois(%)	96,5	**96,8**
Taux de chiffres(%)	92,0	**92,3**

TAB. 5 – *Perplexité et reconnaissance du texte test avec le modèle mixte et le modèle de référence*

Dans le tableau 5, le modèle de référence donne les meilleurs résultats en terme de perplexité, les taux de reconnaissance sur tous les mots sont équivalents. Le modèle mixte améliore la reconnaissance des catégories sémantiques attachées à la tâche de dialogue de 7% pour les jours et de moins de 1% pour les trois autres catégories. L'importance du gain relatif sur la reconnaissance des jours est due aux phénomènes de la catégorie, peu observés dans le texte test. Le faible taux de villes reconnues est dû à une centaine de villes (10 noms différents) qui ne sont pas comprises dans le vocabulaire.

On peut déduire de ces résultats, que si des modèles de langage mixtes ne donnent pas des taux de reconnaissance aussi bons qu'un modèle à base de mots, ils peuvent néanmoins améliorer le système pour des catégories de mots précis qui ont un intérêt spécifique pour la tâche de dialogue.

6 Conclusion

Dans cet article, un classement automatique de mots non étiquetés, fondé sur la similarité des contextes directs de gauche et de droite des mots a été mis en place. Cette classification ne classe pas tous les mots, une partie du vocabulaire reste à l'état de mot. A partir de ce classement, un modèle de langage mixte a été créé et intégré dans le système de reconnaissance du LIMSI pour être évalué en termes de taux de reconnaissance.

La comparaison de ce modèle de langage mixte avec un modèle à base de bigrammes de mots favorise ce dernier autant en taux de reconnaissance qu'en perplexité. Mais l'étude de la reconnaissance de quelques catégories sémantiques importantes pour la tâche de dialogue, montre que les taux obtenus sur des catégories restreintes ne suivent pas l'évolution des taux globaux. L'évaluation du modèle mixte montre qu'il améliore néanmoins la reconnaissance des catégories sémantiques telles que celles des villes, des mois, des jours et des chiffres, catégories clés pour la tâche de dialogue. Tel quel, il devrait augmenter les performances du module de compréhension du système de dialogue, qui utilise principalement ces catégories.

Bibliographie

[Jar96] Jardino M. *Multilingual Stochastic n-gram Class Language Models.* ICASSP'96, Atlanta (USA).

[Jar97] Jardino M., Beaujard C. *Rôle du Contexte dans les Modèles de Langage n-classes, Application et Evaluation sur MASK et RAILTEL.* JST'97, (France).

[Bea97] Beaujard C., Jardino M., Bonneau-Maynard H. *Evaluation of a Class-Based Language Model in a Speech Recognizer.* SPECOM'97, (Roumanie).

[Far96] Farhat A., Isabelle J.F., O'Shaughnessi D. *Clustering probabilistic dependences in stochastic language models based on contextual word similarity.* ICASSP'96, Atlanta (USA).

[Chu90] Church K.W., Hanks P. *Word Association Norms, Mutual Information, and Lexicography.* Computational Linguistics, vol.16, p.22-29, 1990.

[Dag95] Dagan I., Marcus S., Markovitch S. *Contextual Word Similarity and Estimation from Sparse Data.* Computer Speech and Language, vol.9, p.123-152, 1995.

[Huc95] Huckle C.C. *Grouping Words Using statistical Context.* EACL'95, (Irlande).

[Ney94] Ney H., Essen U., Kneser R. *On Structuring probabilistic dependences in stochastic language modelling.* Computer Speech and Language, 1994.

[Lam95] Lamel L.F et al. *Developpment of Spoken Language Corpora for Travel Information.* Euro-Speech'95, (Espagne).

[Pal90] Pallet D.S., Fisher W.M., Fiscus J.G. *Tools for analysis of Benchmark Speech Recognizer* ICASSP'90.

Combinaison de modèles de langages pour l'identification de thèmes

Brigitte Bigi, Renato De Mori, Marc El-Bèze, Thierry Spriet

L. I. A. , Université d'Avignon et des Pays de Vaucluse
CERI-IUP, BP 1228, 84911 Avignon Cedex 9 - France
Tél. : +33 09 90 84 35 00 - Fax : +33 09 90 84 35 01
{brigitte.bigi,renato.demori,marc.elbeze,thierry.spriet}@lia.univ-avignon.fr

Résumé

A new statistical method for Language Modeling and spoken document classification is proposed. It is based on a mixture of topic dependent probabilities. Each topic dependent probability is in turn a mixture of n-gram probabilities and the probability of Kullback-Lieber (KL) distances between key-word unigrams and distribution obtained from the content of a cache memory. Experimental result on topic classification using a corpus of 60 Mwords from the French newspaper *Le Monde* show the excellent performance of the cache memory and its complementary role in providing different statistics for the decision process.

Introduction

La classification en thèmes a suscité de nombreux travaux en recherche documentaire. Plusieurs méthodes ont été développées pour la classification de documents écrits. Récemment, des méthodes statistiques fondées sur les modèles unigrammes ont été proposées ([LIY97]).

L'utilisation de modèles de langage thématiques (ML) fait actuellement l'objet d'un intérêt croissant ([WRI96]; [CAR96]; [PES96]). Pour la transcription automatique de documents de parole ([WRI96]; [CAR96]; [PES96]), les mélanges de ce type de modèles sont souvent employées ([IMA97]; [LIY97]).

En général, un mélange combine les différentes probabilités des ML, en calculant la probabilité d'un mot w_i sachant son histoire h_i de la façon suivante :

$$P(w_i|h_i) = \sum_{j=1}^{J} \lambda_j P_j(w_i|h_i) \qquad (1)$$

où $P_j(w_i|h_i)$ est la probabilité donnée par le j-ème ML et λ_j est un jeu de coefficients satisfaisant la contrainte :

$$\sum_{j=1}^{J} \lambda_j = 1$$

Il est alors possible de rendre λ_j dépendant du thème et égal aux scores qui sont utilisés pour l'identification de thèmes. En fait, $P(w_i|h_i)$ peut être exprimé comme suit :

$$P(w_i|h_i) = \sum_{j=1}^{J} P(w_i, T_j|h_i)$$
$$= \sum_{j=1}^{J} P(w_i|T_j, h_i)P(T_j|h_i) \qquad (2)$$

où T_j est un thème. Le terme $P(w_i|T_j, h_i)$ est évalué avec un modèle n-gram qui dépend du thème. Ceci correspond à faire une approximation telle que h_i est représenté seulement par les $n-1$ mots qui précèdent w_i. Le second terme, $P(T_j|h_i)$, peut être calculé en utilisant une approximation différente pour h_i. Ce dernier peut, en effet, être utilisé pour moduler *dynamiquement* l'importance des éléments du mélange et pour l'identification des thèmes. Cette probabilité peut également être exprimée comme un mélange de probabilités obtenue avec différents modèles, et, plus important encore, elle peut varier dans le temps, au fur et à mesure de l'entrée de nouveaux mots. Ceci rend alors les modèles du langage thématiques **adaptatifs**.

Cet article propose d'utiliser deux modèles. Le premier est composé d'unigrammes thématiques basés sur tous les mots du vocabulaire. Le second modèle repose sur la comparaison entre le contenu d'une mémoire cache et les distributions statiques des mots clés des unigrammes thématiques (avec une valeur constante fixée pour chaque thème et qui est assignée à tous les autres mots qui n'entrent pas dans le cache). Une combinaison linéaire des deux modèles peut être utilisée avec différentes approximations de l'historique des mots :

$$P(T_j|h_i) = \alpha_1 P_1(T_j|W_1^{i-1})$$
$$+\alpha_2 P_2(T_j|d(R_j, cc(i-1))) \qquad (3)$$

où W_1^{i-1} est la séquence $\{w_1, ...w_{i-1}\}$ des premiers $(i-1)$ mots d'un document.

Le premier terme de la formule (3) cumule les informations de l'historique, en contenant tous les mots du vocabulaire. Si ce premier modèle est basé sur les unigrammes, alors :

$$P_1(T_j|W_1^{i-1}) = \prod_{t=1}^{i-1} \frac{P(w_t|T_j)P(T_j)}{\sum_{k=1}^{J} P(w_t|T_k)P(T_k)} \qquad (4)$$

Le problème de ces modèles provient de leur fiabilité limitée et du fait que les mots-clés ont la même importance que ceux qui ne le sont pas.

La caractérisation de thèmes peut être améliorée par l'introduction du second terme de l'équation (3). Alors, un ensemble Γ_j de mots clés est sélectionné pour chaque thème. Sa distribution statistique R_j est obtenue à partir d'un corpus d'apprentissage, et elle est comparée avec la distribution du contenu de $cc(i-1)$ de la mémoire cache, quand le mot w_{i-1} est entré. Le résultat est $d(R_j, cc(i-1))$. C'est l'évaluation de la distance de Kullback-Lieber symétrique qui varie dans le temps, selon l'entrée de nouveaux mots. La probabilité P_2 peut être conditionnée en fonction de la distance entre la distribution du contenu du cache et R_j. Par souci de simplicité, un seul cache est utilisé (un cache différent pour chaque partie de discours pourrait être envisagé).

On suppose que les thèmes forment une partition de l'ensemble des segments. Soit L la taille du vocabulaire. Pour chaque thème, il est possible de trier les L mots selon leur fréquence dans le corpus d'apprentissage et de sélectionner les mots les plus fréquents. Cependant, les mots les plus fréquents qui sont communs à tous les thèmes ne sont pas de bons candidats. Ceux-ci ont été identifiés, et placés dans une liste ; ils ne font pas partie des mots clés. Les premiers $C < L$ mots acceptés de chaque thème sont appelés "*mothèmes*" clés. La comparaison du contenu du cache se limite seulement à ces derniers. Il est alors préférable de normaliser les probabilités des mothèmes clés par rapport aux $L-C$ mots les moins probables auxquels on donne une même probabilité égale à un seuil dans l'unigramme.

Il est possible de faire en sorte que P_2 soit indépendante du thème. En ce sens, la contribution du second terme du mélange de (3) diffère du premier. En fait, plutôt que d'être le produit des probabilités n-grams évaluées sur tout le vocabulaire, il est la probabilité d'une distance évaluée avec des distributions limitées aux mots clés des thèmes.

1. Probabilités dépendantes du thème

Soit $P_{jC} = \sum_{i=1}^{C} P_1(w_i|T_j)$, la somme des probabilités statiques des unigrammes dans le thème T_j. Pour le calcul de la distance, les probabilités de ces C mots sont multipliées par une constante γ_j dépendante du thème. Alors, la probabilité seuil des autres $L-C$ mots est :

$$P_{jr}(w) = \frac{1-\gamma_j P_{jC}}{L-C}$$

Après expérimentation, C est fixé à 4 000 pour chacun des thèmes. De meilleurs choix des C mots clés pour chaque thème restent possibles. Pour peu qu'un étiquetage soit disponible, on peut par exemple choisir de ne retenir que les mots de certaines classes syntaxiques.

En supposant qu'il y ait m mots dans le cache, et que ces mots correspondent à G mots différents, la probabilité d'un mot dans le cache est donnée par :

$$P_{cache}(w) = \frac{n(w)+\beta}{\beta L + m} \quad (5)$$

où $n(w_i)$ est le nombre d'occurrences du mot w_i dans le cache, β est un seuil constant et $\sum n(w) = m$. La probabilité des mots absents du cache est :

$$P_{fC}(w) = \frac{\beta}{\beta L + m} \quad (6)$$

Ainsi, pour avoir une probabilité seuil identique pour les mots absents du cache et ceux des modèles thématiques normalisés, il est suffisant de poser :

$$P_{jr}(w) = P_{fC}(w)$$

$$\frac{1-\gamma_j P_{jC}}{L-C} = \frac{\beta}{\beta L + m}$$

Ce qui donne :

$$\gamma_j = \{1 - \beta\frac{L-C}{\beta L+m}\}\frac{1}{P_{jC}} = \frac{1}{P_{jC}}(1-A) \quad (7)$$

La probabilité normalisée pour un mot dans le thème T_j devient :

$$P'(w_k|T_j) = \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)}(1-A)$$

$$= (1-A)f'(w_k|T_j) \quad (8)$$

avec $f'(w_k|T_j) = \frac{n_j(w_k)}{\sum_{i=1}^{C} n_j(w_i)}$.

où $n_j(w_i)$ est le compte des mots dans le thème T_j.

2. Évaluation de la distance

L'évaluation de la distance commence quand le cache contient assez de mots. $d_j(n)$ est la distance relative au thème j après avoir lu les n premiers mots. $P_{cache}(n, w_i)$ est $P_{cache}(w_i)$ à l'insertion dans le cache du n-ème mot du document en cours de traitement. Tous les mots qui ne sont des mots clés dans aucun des thèmes ont la probabilité seuil dans la distribution dépendante du thème et dans le cache. Ainsi, leur contribution à la distance de KL est nulle. Tous ces mots peuvent être ignorés dans le calcul du second terme de la formule (3). La distance de KL symétrique peut alors être calculée comme suit :

$$d_j(n) = \sum_{i=1}^{C} (P_{cache}(n, w_i) - Bf'(w_i|T_j))$$

$$log\left(\frac{P_{cache}(n, w_i)}{Bf'(w_i|T_j)}\right) \quad (9)$$

Un mot qui n'est mot clé pour aucun thème et qui apparaît dans le cache fournirait une contribution égale dans l'évaluation des distances de chacun des thèmes.

Pour cette raison, ces mots n'ont pas besoin d'entrer dans le cache.

Les coefficients α_1 et α_2 de (3) doivent sommer à 1. Ils sont déterminés par l'interpolation proposée en ([KUH90]). Le coefficient α_2 peut être pris plus petit si le cache n'est pas plein. Il faut noter qu'il est possible que le cache ne contienne aucun des mots clés d'un thème. Si une probabilité nulle était affectée au cache pour chacun des mots clés, la distance aurait une valeur infinie. L'introduction d'une probabilité seuil évite cette situation.

La contribution de la distance de KL pour l'évaluation de la probabilité d'un mot peut être de différents types. Considérons d'abord le vecteur D composé des distances entre le cache et chacun des thèmes :

$$D(n) = \lfloor d_1(n), ..., d_j(n), ...d_J(n) \rfloor$$

La probabilité suivante peut être considérée :

$$P_2(T_j|D(n)) = \frac{P\lfloor D(n)|T_j \rfloor P(T_j)}{\sum_{k=1}^{J} P\lfloor D(n)|T_k \rfloor P(T_k)} \qquad (10)$$

Différentes approximations pour l'évaluation des probabilités antérieures sont possibles. Les expériences avec les probabilités des distances n'ont pas encore été réalisées.

Tab. 1: Thèmes et leurs codes

1	Etranger	5	Affaire, Economie
2	Histoire	6	Culture, Arts, Livres, Media
3	Sciences	7	Politique
4	Sports		

3. Résultats

Trois années des articles du journal *"Le Monde"* ont été utilisées, ce qui représente un total de 60 millions de mots pour un vocabulaire de 500 000 mots. Les thèmes de ces articles ne sont pas connus, mais, grossièrement, les secteurs de rédactions du journal ont été pris comme thèmes. Sept d'entre-eux en ressortent ; leurs intitulés et codes sont définis dans la table 1. Un corpus de test de 1 021 paragraphes pris parmi les articles a été extrait, et n'a pas été utilisé dans l'apprentissage.

Des résultats préliminaires sont présentés dans les tables 2 et 3. U représente le thème retenu par l'unigramme seulement (la probabilité P_1 de la formule (4)), C représente le thème retenu par le modèle cache (dans notre cas c'est une simple normalisation de la distance de KL), S représente le thème d'origine du paragraphe dans le journal, C out représente le cas le cache contient moins de 5 mots clés. Les probabilités sont évaluées mot à mot et les résultats présents sont basés sur les valeurs au dernier mot de chaque paragraphe. Le symbole = représente l'accord, $\neq$ représente le désaccord,

La table 3 résume tous les résultats de l'identification thématique du jeu de test en fonction des différentes stratégies utilisées. La première ligne réfère au cas où le cache et l'unigramme sont en accord avec le label. Dans la deuxième ligne, on ajoute le cas où l'unigramme est en accord avec le label et le cache ne contient pas assez de données pour être pris en compte. La troisième ligne intègre à la deuxième, la décision prise par l'unigramme seulement. La quatrième ligne intègre à la deuxième, la décision prise par le cache seulement, excepté quand le cache ne contient pas assez de mots. La dernière ligne correspond à la décision alterne entre le cache et l'unigramme. La stratégie de combinaison consiste à utiliser le cache pour décision finale seulement s'il contient assez de mots et si la différence entre les scores du cache des premiers et seconds candidats est supérieure à un seuil. Ce dernier est déterminé durant l'apprentissage. Ces résultats préliminaires avec une règle grossière de décision montrent une augmentation des performances grâce à l'utilisation du cache. Avec la contribution du cache, une précision supérieure à 78 % est obtenue.

Conclusion

Les expériences décrites dans cet article montrent les avantages de l'utilisation conjointe d'un modèle cache et d'un modèle unigramme. Leur complémentarité permet une astucieuse combinaison et peut aboutir à une amélioration substantielle de la classification thématique. La tâche de la classification thématique est très délicate à traiter Elle repose en effet sur la décision humaine et son caractère subjectif fait qu'elle est souvent remise en cause par d'autres personnes. Les recherches futures s'orientent vers plusieurs directions. L'emploi d'un lemmatiseur servira à pallier les lacunes du cache en réduisant de fait la taille du vocabulaire. Différents types de probabilités pour les distances peuvent être étudiés étant donné leur utilisation dans les ML adaptatifs pour la reconnaissance automatique de la parole (ASR). L'utilisation de plusieurs caches peut être introduite. Les mots pourront être répartis selon leur classe sémantique connue grâce à l'utilisation d'un programme d'étiquetage sémantique. D'autres utilisations des mots et de leurs classes sémantiques peuvent aussi être envisagées.

Références

[CAR96] B. A. Carlson. Unsupervised topic clustering of switchboard speech messages. Proc. of the *IEEE International Conference on Acoustics*, pages 315–319, Atlanta GA, 1996.

[IMA97] T. Imai R. Schwartz F. Kubala and L. Nguyen. Improved Topic Discrimination of Broadcast News Using a Model of Multiple Simultaneous Topics. Proc. of the *IEEE International Conference on Acoustics*, Speech and Signal Processing, pages 727–730. Munich, Germany, 1997.

[KUH90] R. Kuhn R. De Mori. A cache-based na-

Tab. 2: Résultats préliminaires de classification thématique, en nombre de segments

Thème	U ≠ S C ≠ S U ≠ C	U ≠ S C ≠ S U = C	U = S C ≠ S U ≠ C	U ≠ S C = S U ≠ C	U = S C = S U = C	U ≠ S C out	U = S C out	Total
1	4	10	29	13	75	2	6	139
2	0	0	1	0	6	0	0	7
3	19	4	2	17	48	1	0	91
4	15	7	12	57	21	0	2	114
5	2	3	16	6	129	0	3	159
6	15	21	55	24	244	6	16	381
7	15	21	18	18	51	3	4	130
Total	70	66	133	135	*574*	12	31	1021

Tab. 3: Résultats avec stratégie de combinaison

Stratégie	N	%
U = C = S	574	56,22
+ C out	605	59,26
Deux premiers + Unigramme	738	72,28
Deux premiers + Cache	740	72,48
Stratégie de combinaison	800	78,35

tural language model for speech recognition. In *IEEE Trans. Pattern anal. Machine Intell*, PAMI-12(6), pages 570–582, 1990.

[LIY97] H. Li and K. Yamamishi. Document classification using a finite mixture model. Proc. of the *Conference of the Association for Computational Linguistics*, pages 39–47, Madrid, Spain, 1997.

[PES96] B. Peskin S. Conolly L. Gillick S. Lowe D. McAllaster V. Nagesha P. Van Mulbregt S. Wegmann Improvement in SWITCHBOARD recognition and topic identification. Proc. of the *IEEE International Conference on Acoustic*, Speech and Signal Processing, pages 303–306, Atlanta GA, 1996.

[SPR97] T. Spriet M. El-Bèze. Introduction of Rules into a Stochastic Approach for Language Modeling In NATO ASI series F, Edited by K. M. Ponting, Sprnger Verlag, Berlin New-York, 1997.

[WRI96] J. H. Wright M. J .Carry and E. S. Parris Statistical models for topic identification using phoneme substrings Proc. of the *IEEE International Conference On Acoustics*, Speech and Signal Processing, pages 307–310, Atlanta GA, 1996.

Apport de l'utilisation de classes de mots pour la définition d'un nouveau lissage de type absolute discounting

G. Damnati & J. Simonin

France Télécom CNET/DIH/DIPS
2 av. P. Marzin, 22307 Lannion Cedex, France
Tél.: (33) 2 96 05 13 88 - Fax: (33) 2 96 05 35 30
e-mail: Geraldine.Damnati@cnet.francetelecom.fr

ABSTRACT

This paper describes a new smoothing technique, derived from the absolute discounting smoothing approach, which applies efficiently to a word bigram language model. The originality of this smoothing technique lies in the fact that information from a class-based model is taken into account, as an intermediate level in the classical absolute discounting hierarchical scheme. Adding in this class-based intermediate level leads to improved performance in terms of perplexity and recognition error rate, as compared to classical smoothing techniques such as threshold smoothing, linear discounting or absolute discounting smoothing. The so called "absolute discounting + classes" smoothing proposed here, is evaluated through a speech recognition system on a spontaneous speech corpus composed of human-computer spoken dialogues.

1. INTRODUCTION

Pour un processus de reconnaissance n'utilisant qu'une seule passe, et donc un seul modèle de langage, celui-ci est le plus fréquemment de type bigramme. Dans ces conditions restrictives, le choix d'un lissage est crucial pour améliorer les performances. Nous proposons ici un nouveau type de lissage pour lequel les probabilités des événements non vus dans le corpus d'apprentissage (des paires de mots non vues dans le cas du bigramme), sont estimées à partir d'une classification des mots du vocabulaire.

Le lissage par absolute discounting [Ney94] est l'un des lissages les plus performants pour les bigrammes de mots. Le lissage proposé dans cet article utilise, en complément à un lissage de type absolute discounting, un bigramme de classes qui permet d'estimer les probabilités des événements non vus. Les informations liées aux classes sont utilisées lorsqu'une paire de mots n'est pas observée dans le corpus d'apprentissage, auquel cas l'événement considéré devient la paire de classes correspondante Du fait d'une couverture plus large des événements possibles, le modèle à base de classes offre une meilleure estimation des non vus que le simple modèle d'absolute discounting, qui ne bénéficie pour estimer ceux-ci que d'une information de type unigramme.

Afin d'obtenir une classification des mots du vocabulaire, nous utilisons une méthode de classification ascendante hiérarchique [Bro92], basée sur la notion de voisins réciproques où plusieurs paires de classes peuvent être regroupées dans une même itération suivant le critère de l'information mutuelle [Dam97].

Dans une première partie, nous présentons la description de ce lissage de type absolute discounting amélioré désigné sous le nom de "absolute discounting + classes". La deuxième partie regroupe l'évaluation comparée de différents lissages de bigrammes de mots, tels que le linear discounting, l'absolute discounting et l'"absolute discounting + classes" proposé ici, ainsi qu'un modèle d'interpolation d'un bigramme de mots et d'un bigramme de classes. Ces évaluations sont faites en termes de perplexité et de taux d'erreurs mots à l'issue d'un processus de reconnaissance de la parole continue.

2. LISSAGE "ABSOLUTE DISCOUNTING + CLASSES"

Le lissage est un moyen classique d'améliorer les performances d'un système de reconnaissance de la parole continue. Il permet en effet de traiter le problème des événements non vus lors de l'estimation du modèle de langage.

2.1 Lissage par Absolute Discounting

Le lissage par absolute discounting est un lissage de type back-off [Kat87] dont le principe est de redistribuer parmi les événements non vus une masse de probabilité prélevée par ailleurs sur chacun des événements vus.

Dans le cas d'un bigramme de mots, si N_m (m) représente le nombre d'occurrences, dans le corpus d'apprentissage, du mot m et N_m (m_0,m_1), celui de la paire de mots (m_0,m_1),la formule s'écrit de la façon suivante :

$$\tilde{P}_a(m_1/m_0)=\begin{cases}\dfrac{N_m(m_0,m_1)-d}{N_m(m_0)}, & \text{si } N_m(m_0,m_1)>0\\ \gamma(m_0)\dfrac{N_m(m_1)}{\nu(m_0)}, & \text{si } N_m(m_0,m_1)=0,\ N_m(m_0)>0\\ \tilde{P}_u(m_1), & \text{si } N_m(m_0)=0\end{cases}$$

avec $\gamma(m_0) = \frac{1}{N_m(m_0)} \sum_{m_i,\ N_m(m_0,m_i)>0} d$

et $\nu(m_0) = \sum_{m_i,\ N_m(m_0,m_i)=0} N_m(m_i)$

$\tilde{P}_u(m)$ représente la probabilité de l'unigramme m.

Le paramètre d utilisé dans la formule de l'absolute discounting est fixé pour toutes les paires de mots observées, indépendamment de leur nombre d'occurrences dans le corpus d'apprentissage.

2.2 Modèles à base de classes

Nous nous plaçons ici dans le cadre d'un modèle de bigramme à base de classe, pour lequel un mot ne peut appartenir qu'à une seule classe. Pour une telle modélisation, les événements observés dans le corpus d'apprentissage ne sont plus des paires de mots mais des paires de classes. Ainsi, étant donnée une partition C du vocabulaire, la probabilité d'un mot m_1 de classe $C(m_1)$ sachant m_0 de classe $C(m_0)$ est estimée de la façon suivante :

$$\tilde{P}_c(m_1/m_0) = \tilde{P}_1(m_1/C(m_1))\,\tilde{P}_2(C(m_1)/C(m_0)) \quad (1)$$

Si $N_m(m)$ représente le nombre d'occurrences du mot m, $N_m(m_0,m_1)$, le nombre d'occurrences de la paire de mots (m_0,m_1), $N_c(m)$, le nombre d'occurrences de la classe C(m) et, $N_c(m_0,m_1)$, le nombre d'occurrences de la paire de classes $(C(m_0),C(m_1))$ dans le corpus utilisé lors de l'apprentissage du bigramme, la formule (1) peut se réécrire en ne considérant que les fréquences d'apparition :

$$\tilde{P}_c(m_1/m_0) = \frac{N_m(m_1)}{N_c(m_1)}\,\frac{N_c(m_0,m_1)}{N_c(m_0)}$$

L'intérêt des classes est de fournir une sorte de lissage en modifiant les observations. Le recouvrement des événements vus est naturellement plus conséquent avec une observation basée sur les classes que sur les mots. Ainsi, plus le nombre de classes considéré est réduit et plus la proportion d'événements observés est grande mais, étant donnée la formule (1), moins l'estimation des probabilités correspondant à ces événements est précise. Un compromis doit être trouvé entre couverture des événements et qualité de l'estimation.

2.3 Nouvelle estimation du bigramme

Le problème de l'estimation des probabilités des événements non vus peut aussi être envisagé de la façon suivante : utiliser une source d'information plus générale (i.e. qui offre une plus grande couverture des événements possibles), pour estimer les probabilités des événements non observés dans la source d'information courante.

La deuxième source d'information peut être par exemple un second corpus. Dans un contexte de dialogue, le premier corpus serait plus spécifique à l'application ciblée et le deuxième plus général. L'utilisation d'un second corpus permet, avec un lissage de structure hiérarchique de type back-off [Kat87], d'attribuer à un événement non vu dans le corpus spécifique une probabilité estimée à l'aide du second corpus [Bes95].
Nous nous proposons ici d'exploiter cette approche en utilisant non pas un deuxième corpus pour estimer un événement non vu dans le corpus d'apprentissage, mais un modèle offrant un recouvrement plus large des événements, à savoir un modèle à base de classes.

Le lissage "absolute discounting + classes" proposé se définit alors comme suit.

$$\tilde{P}(m_1/m_0) = \begin{cases} \dfrac{N_m(m_0,m_1) - d}{N_m(m_0)} & \text{, si } N_m(m_0,m_1)>0 \\ \alpha_1(m_0)\,\tilde{P}_c(m_1/m_0), & \text{si } N_m(m_0,m_1)=0 \text{ et } N_c(m_0,m_1)>0 \\ \alpha_2(m_0)\,\tilde{P}_u(m_1) & \text{, si } N_m(m_0,m_1)=0,\ N_c(m_0,m_1)=0 \text{ et } N_m(m_0)>0 \\ \tilde{P}_u(m_1) & \text{, si } N_m(m_0)=0 \end{cases}$$

Les paramètres utilisés lors de ce lissage sont définis de la façon suivante :

$$-\ \alpha_1(m_0) = \frac{\alpha\,\gamma(m_0)}{\sum_{\substack{m_i,\ N_m(m_0,m_i)=0 \\ N_c(m_0,m_i)>0}} \tilde{P}_c(m_i/m_0)}$$

$$-\ \alpha_2(m_0) = \frac{(1-\alpha)\,\gamma(m_0)}{\sum_{\substack{m_i \\ N_c(m_0,m_i)=0}} \tilde{P}_u(m_i)}$$

avec $\gamma(m_0)$ et $\tilde{P}_u(m)$ définis comme au paragraphe 2.1
Les paramètres nécessaires pour la création de ce bigramme sont alors le classique paramètre *d* d'absolute discounting, et le paramètre α, $\alpha \in]0,1[$. Suite à de nombreuses expérimentations, d optimisé vaut, de manière classique, 0,7 et α vaut 0,5.

3. EVALUATIONS

Les évaluations du lissage proposé précédemment sont réalisées sur un corpus de parole spontanée issu de dialogues homme-machine sur le serveur en démonstration AGS [Sad96].

3.1 Description du corpus

Ce corpus est enregistré à travers le réseau téléphonique. Le vocabulaire de l'application de dialogue AGS comprend 876 mots. Le corpus d'apprentissage du bigramme est constitué de 9850 phrases pour 49610 occurrences de 822 mots du lexique. Le corpus de test contient 724 phrases pour 3584 mots. Le système de reconnaissance de la parole continue utilisé est indépendant du locuteur.

3.2 Comparaison statistique des lissages "absolute discounting + classes" et absolute discounting

Avant d'évaluer les performances des différents lissages nous effectuons quelques observations d'ordre statistique concernant l'impact des lissages sur la répartition des probabilités.
La classification utilisée ici a été optimisée au préalable. Il s'avère qu'un nombre de classes égal à 100 permet d'obtenir les meilleures performances en termes de perplexité. Le nombre de classes utilisées est relativement faible; ceci est directement lié à la notion de couverture des événements vus. En effet, moins le nombre de classes est élevé et plus le nombre de paires de mots vérifiant $N_c(m_0,m_1)>0$ est élevé, or la couverture du modèle à base de classes doit être suffisamment importante pour que le niveau ajouté à l'absolute discounting ait un apport significatif.

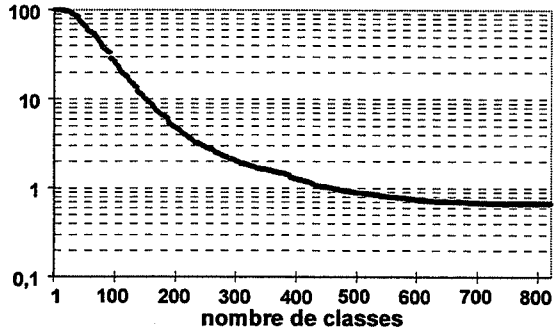


Figure 1 : Couverture des événements observables par le modèle à base de classes obtenu par classification hiérarchique.

Avec un modèle à base de mots, sur les 822^2 événements observables[1], seuls 0,57% sont représentés dans le corpus d'apprentissage, c'est-à-dire que 0,57% des événements sont concernés par le premier niveau de la formule de l'absolute discounting ($N_m(m_0,m_1)>0$). Le modèle à 100 classes offre pour sa part une couverture de 15,4% des événements observables, c'est à dire concernés par les deux premiers niveaux de la formule de l'"absolute discounting + classe". Par ailleurs, 70,81 % des paires de mots rencontrées dans l'ensemble de test sont observées dans l'apprentissage, et le modèle à base de 100 classes offre une couverture de 80,40 % des événements du corpus de test.

La figure 2 représente sous forme d'histogramme la répartition des valeurs des 876^2 probabilités possibles pour les deux lissages considérés. L'effet de l'ajout d'un niveau intermédiaire utilisant un modèle à 100 classes se traduit par une répartition plus équilibrée des valeurs des probabilités. Le pic présent autour de la valeur 10^{-5} est estompé par la prise en compte des classes dans un niveau intermédiaire. En fait, les valeurs attribuées aux événements non observés dans le corpus d'apprentissage le sont de façon plus dynamique, ce qui est plus satisfaisant si l'on considère le caractère discriminant du modèle de langage.

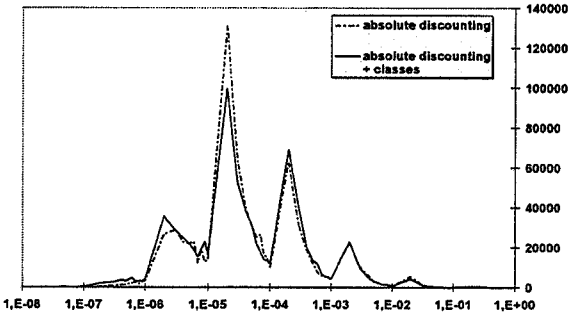


Figure 2 : Histogramme des valeurs des probabilités.

3.2 Evaluation des différents lissages

L'évaluation du lissage "absolute discounting + classes" est faite à partir du calcul de la perplexité et du taux d'erreurs mots obtenu lors de la reconnaissance. Une comparaison est établie entre le lissage proposé et des lissages classiques, tels que le lissage par seuil, le lissage de type linear discounting et celui de type absolute discounting. Par ailleurs, afin d'établir une comparaison avec un modèle de langage combinant lui aussi des informations liées aux mots et aux classes, nous comparons le lissage proposé à un modèle issu d'une interpolation linéaire [Jel80] d'un bigramme de mots lissé par absolute discounting et d'un bigramme de 100 classes[2].

Table 1 : Comparaison de la perplexité sur le corpus de test pour différents lissages du bigramme de mots.

Type de Lissage	Perplexité
Seuil	21,81
Linear Discounting	22,06
Absolute Discounting	20,84
Interpolation linéaire	19,74
"Absolute Discounting + Classes"	19,88

Dans le tableau 1, une première série de résultats, concernant la perplexité sur le corpus de test, met en évidence l'apport de l'information liée aux classes dans un lissage de bigram de mots. Les deux lissages "absolute discounting + classes" et interpolation linéaire offrent une perplexité équivalente sur l'ensemble de test.

[1] Par événements observables, nous entendons événement impliquant des mots d'occurrence non nulle dans le corpus d'apprentissage.

[2] L'interpolation linéaire s'écrit sous la forme suivante :

$\tilde{P}(m_1/m_0) = \lambda \tilde{P}_a(m_1/m_0) + (1 - \lambda) \tilde{P}_c(m_1/m_0)$ avec λ=0.8

Afin d'avoir une comparaison plus large entre ces différents lissages, nous ne comparons pas seulement leurs performances en termes de taux d'erreurs mais nous étudions l'évolution de leurs performances en faisant varier le paramètre d'élagage en fin de mots utilisée durant le décodage de Viterbi [Ste94]. La signification d'une variation à la hausse de ce paramètre est une augmentation du nombre de successeurs possibles au mot décodé et donc une augmentation du pourcentage de l'espace de recherche activé lors du décodage. En effet, plus la valeur de l'heuristique est élevée et plus il sera accepté parmi les successeurs possibles d'un mot m_0, de mots m pour lesquels la probabilité bigramme correspondante $P(m|m_0)$ est faible. Ainsi, l'évolution des performances d'un lissage en fonction du paramètre d'élagage en fin de mot, est un bon indicateur du comportement de ce lissage lorsque l'on autorise au cours du décodage des événements de faible probabilité.

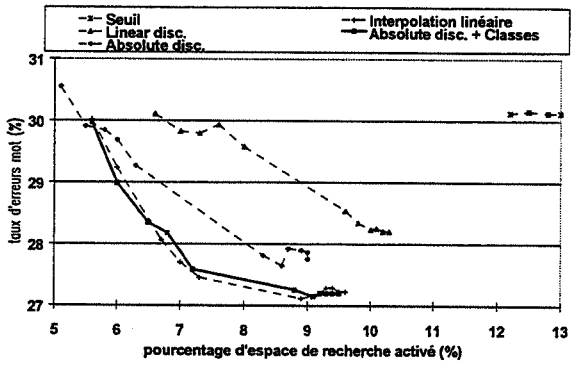


Figure 3 : Comparaison des taux d'erreurs mots pour le corpus AGS en fonction du pourcentage d'espace de recherche activé en moyenne au cours du décodage.

Cette série d'évaluations confirme l'efficacité du lissage "absolute discounting + classes" par rapport aux lissages classiques n'utilisant pas d'information liée aux classes.

Le lissage par seuil n'est pas affecté par l'évolution du paramètre d'élagage en fin de mots dans la mesure où les valeurs des probabilités sont supérieures à un seuil qui est déjà inclus parmi les possibles dans le cas de la plus faible valeur du paramètre d'élagage. Par contre, les autres lissages offrent une répartition des valeurs des probabilités plus étalée et l'évolution de l'élagage a une influence sur l'ouverture de l'espace de recherche et donc sur les performances. L'étude des courbes de la figure 3 montre que les lissages utilisant l'information liée aux classes ont un meilleur comportement lorsque l'espace de recherche s'ouvre (i.e. lorsque l'on élargit les possibilités lors du décodage). A la valeur minimale du paramètre d'élagage considéré ici, les lissages ont tous des performances comparables, puis, lorsque ce paramètre augmente, l'amélioration est plus sensible pour les lissages "absolute discounting + classe" et interpolation linéaire. Ceci signifie que les faibles valeurs prises en compte lorsque le pourcentage d'espace de recherche activé est plus élevé correspondent à une meilleure estimation des événements non vus. En d'autres termes, il est possible de rattraper un plus grand nombre d'erreurs en acceptant des événements peu probables (et donc en augmentant le nombre de successeurs possibles en moyenne) dans le cas des lissages utilisant une information liée aux classes.

4. CONCLUSION

Nous avons montré dans cette étude qu'un lissage appliqué à un bigramme de mots pouvait être enrichi par une information issue de classes construites de façon automatique. Par ailleurs le lissage "absolute discounting + classes" proposé ici intègre cette information de façon hiérarchique dans un schéma de type back-off. Cette structure hiérarchique présente l'intérêt de pouvoir intégrer des informations d'origines variées comme par exemple différents niveaux de recouvrement pour une même classification, différents types de classifications (automatiques, syntaxiques...).

BIBLIOGRAPHIE

[Ney94] : H. Ney, U. Essen, R. Knesser (1993), "On Structuring Probabilistic Dependencies in Stochastic Language Modelling", *Computer Speech and Language*, Vol. 7, pages 101-138.

[Bro92] : P.F. Brown, V.J. Della Pietra, P.V. de Souza, J.C. Lai, R.L. Mercer (1992), "Class-Based n-gram Models of Natural Language", *Computational Linguistics*, 18(4), pages 467-479.

[Dam97] : G. Damnati and J. Simonin (1997), "A Novel Tree-Based Clustering Algorithm for Statistical Language Modeling", *Proceedings of Eurospeech Conference*, Rhodes, Grèce, pages 2715-2718.

[Kat87] : S. Katz (1987), "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Transactions on ASSP*, Mars 1987, 35(3) : pages 400-401.

[Bes95] : S. Besling, H.G. Meier (1995), "Language Model Speaker Adaptation Modeling", *Proceedings of Eurospeech Conference*, Madrid, Espagne, pages 1755-1758.

[Ste94] : V. Steinbiss, B Tran and H. Ney (1994), "Improvements in Beam Search", *Proceedings of ICSLP*, Yokohama, Japon, pages 2143-2146.

[Sad96] : D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget and J. Simonin "Effective Human-Computer Cooperative Spoken Dialogue : the AGS Demonstrator". *Proceedings of ICSLP*, Philadelphia, USA, Octobre 1996, pages 546-549.

[Jel80] : F. Jelinek and R. Mercer 1980, "Interpolated Estimation of Markov Source Parameters from Sparse Data", *Pattern Recognition in Practice*, Ed. Gelsema et Kanal, pages 381-397.

Modélisation statistique du langage avec un modèle bi-multigramme

Sabine Deligne *Yoshinori Sagisaka*

ATR-ITL, dept1, 2-2 Hikaridai
Seika cho, Soraku gun, Kyoto fu 619-0224, Japan.
sdeligne@itl.atr.co.jp sagisaka@itl.atr.co.jp

Abstract

In this paper, we present a stochastic language modeling tool which aims at retrieving variable-length phrases (multigrams), assuming n-gram correlations between them. The phrase retrieval can be intermixed with a phrase clustering procedure, so that data are iteratively structured at both a paradigmatic and syntagmatic level. Perplexity results on ATR travel arrangement data with a bi-multigram model (assuming bigram correlations between the phrases) come very close to the trigram scores with a reduced number of entries in the language model. Also the ability of the class version of the model to merge semantically related phrases into a common class is illustrated.

1. Introduction

Dans cet article, nous présentons un modèle de langage statistique dérivé du modèle multigramme introduit dans [1]. Il s'agit d'un modèle stochastique à base de séquences de mots de longueurs variables. La formulation théorique proposée dans [4] supposait l'indépendance des séquences. Dans la section 2, cette hypothèse est levée au profit de dépendances de type n-gramme entre les séquences [5], et un aspect paradigmatique est intégré au modèle, qui permet la classification des séquences. Dans la section 3, cette approche est évaluée sur une base de données d'informations touristiques en termes de perplexité et de taille du modèle.

2. Formulation théorique

2.1. *Distribution multigramme*

Dans l'approche multigramme, l'hypothèse est faite que les phrases s'articulent en séquences de mots de longueurs variables, appelées multigrammes. La vraisemblance d'une phrase s'obtient en sommant les vraisemblances de toutes les segmentations possibles, le calcul de la vraisemblance d'une segmentation particulière étant déterminé par le modèle décrivant les dépendances entre les multigrammes. Nous appelons bi-multigramme le modèle où des dépendances bigrammes sont supposées entre les multigrammes. En limitant à 3 mots la taille maximale d'une séquence, la vraisemblance bi-multigramme de "a b c d" s'écrit:

$$\sum \begin{cases} p([a] \mid \#)\, p([b] \mid [a])\, p([c] \mid [b])\, p([d] \mid [c]) \\ p([a] \mid \#)\, p([b] \mid [a])\, p([cd] \mid [b]) \\ p([a] \mid \#)\, p([bc] \mid [a])\, p([d] \mid [bc]) \\ p([a] \mid \#)\, p([bcd] \mid [a]) \\ p([ab] \mid \#)\, p([c] \mid [ab])\, p([d] \mid [c]) \\ p([ab] \mid \#)\, p([cd] \mid [ab]) \\ p([abc] \mid \#)\, p([d] \mid [abc]) \end{cases}$$

Pour la présentation du formalisme du modèle dans cette section, on suppose, plus généralement, des dépendances $\overline{n}$-grammes entre les séquences, et on note n la taille maximale d'une séquence (dans l'exemple précédent, $\overline{n}$=2 et n=3). Soit un corpus W et $\{S\}$ l'ensemble des segmentations possibles sur ce corpus ; la vraisemblance de W s'écrit:

$$\mathcal{L}(W) = \sum_{S \in \{S\}} \mathcal{L}(W, S) \qquad (1)$$

et celle d'une segmentation S de W:

$$\mathcal{L}(W, S) = \prod_{\tau} p(s_{(\tau)} \mid s_{(\tau-\overline{n}+1)} \cdots s_{(\tau-1)}) \qquad (2)$$

avec $s_{(\tau)}$ séquence de rang (τ) dans la segmentation S. Le modèle est donc complètement spécifié par l'ensemble des probabilités $\overline{n}$-grammes associé à l'ensemble $\{s_i\}_i$ des séquences distinctes pouvant être formées en combinant 1, 2, ... ou n mots du vocabulaire. L'estimation au maximum de vraisemblance de ces probabilités peut se formuler comme un problème d'estimation à partir de données incomplètes [6], la donnée cachée étant dans ce cas la segmentation S du corpus. Il vient alors la formule d'estimation itérative suivante: $p^{(k+1)}(s_{i_{\overline{n}}} \mid s_{i_1} \ldots s_{i_{\overline{n}-1}}) =$

$$\frac{\sum_{S \in \{S\}} c(s_{i_1} \ldots s_{i_{\overline{n}-1}} s_{i_{\overline{n}}},\, S) \times \mathcal{L}^{(k)}(S \mid W)}{\sum_{S \in \{S\}} c(s_{i_1} \ldots s_{i_{\overline{n}-1}},\, S) \times \mathcal{L}^{(k)}(S \mid W)} \qquad (3)$$

où $\mathcal{L}^{(k)}(S|W)$ est la vraisemblance conditionnelle de la segmentation S à l'itération (k), et $c(s_{i_1} \ldots s_{i_{\overline{n}}},\, S)$ le nombre d'occurences du $\overline{n}$-uplet de séquences $s_{i_1} \ldots s_{i_{\overline{n}}}$ dans S. La formule (3) peut être implémentée au moyen d'un algorithme forward-backward avec une complexité $O(n^{\overline{n}}T)$ pour un corpus de T mots (cf. l'algorithme présenté en annexe pour $\overline{n} = 2$). En approximant, à chaque itération (k), la vraisemblance du corpus avec celle de la segmentation la plus probable $S^{*(k)}$, la formule de réestimation se simplifie en:

$$p^{(k+1)}(s_{i_{\overline{n}}} | s_{i_1} \ldots s_{i_{\overline{n}-1}}) = \frac{c(s_{i_1} \ldots s_{i_{\overline{n}-1}} s_{i_{\overline{n}}}, S^{*(k)})}{c(s_{i_1} \ldots s_{i_{\overline{n}-1}}, S^{*(k)})} \qquad (4)$$

$S^{*(k)}$ étant déterminé par un algorithme de Viterbi.

2.2. *Classes de bi-multigrammes*

Les modèles à base de classes de séquences reposent en général sur une étape préalable de classification des mots [8]: une classe de séquences est définie pour chaque combinaison distincte de classes de mots observée. En conséquence, seules des séquences ayant la même longueur peuvent appartenir à une même classe [1]. Nous proposons de lever cette limitation en classant directement des séquences, au lieu de mots. A cette fin, la procédure d'apprentissage bi-multigramme ($\overline{n} = 2$) de la section 2.1 peut être modifiée de sorte que chaque itération comprenne 2 étapes:

• 1 Classification des multigrammes:

$$\{ p^{(k)}(s_j \mid s_i) \}$$

$$\longrightarrow \quad \{p^{(k)}(C_{q(s_j)} \mid C_{q(s_i)}),\ p^{(k)}(s_j \mid C_{q(s_j)}) \}$$

• 2 Réestimation bi-multigramme:

$$\{ p^{(k)}(C_{q(s_j)} \mid C_{q(s_i)}),\ p^{(k)}(s_j \mid C_{q(s_j)}) \}$$

$$\longrightarrow \quad \{p^{(k+1)}(s_j \mid s_i) \}$$

L'étape 1 consiste à classer automatiquement les séquences, connaissant leur distribution bi-multigramme. Dans nos expériences, cette étape de classification est réalisée de façon à maximiser l'information mutuelle moyenne entre séquences adjacentes [2]. L'étape 2 consiste à réestimer la distribution bi-multigramme des séquences d'après la formule (3) ou (4), la vraisemblance $\mathcal{L}^{(k)}(W,S)$ étant calculée avec l'hypothèse de classes, comme:

$$\prod_{\tau} p^{(k)}(C_{q(s_{(\tau)})} \mid C_{q(s_{(\tau-1)})})\, p^{(k)}(s_{(\tau)} \mid C_{q(s_{(\tau)})}) \quad (5)$$

L'étape 1 garantit que la classification des séquences est optimale (au sens de l'information mutuelle moyenne) par rapport à la distribution bi-multigramme estimée à l'itération précédente ; et l'étape 2 garantit que la distribution bi-multigramme optimise la vraisemblance du corpus étant donnée la nouvelle classification. Le corpus est ainsi structuré itérativement, selon une dimension à la fois paradigmatique (étape 1) et syntagmatique (étape2).

3. Expériences

3.1. *Protocole et base de données*

Protocole d'évaluation Une motivation pour l'estimation de probabilités bi-multigrammes est d'améliorer la capacité de prédiction d'un modèle 2-gramme de mots, tout en maintenant le nombre de paramètres du modèle en-deçà de celui d'un modèle 3-gramme (réduisant de la sorte la complexité de l'algorithme de recherche lors du décodage par un système de reconnaissance). Le pouvoir de prédiction d'un modèle de langage est usuellement mesuré par la valeur de perplexité:

$$PP = e^{-\frac{1}{T}\log \mathcal{L}(W)}$$

avec T le nombre de mots dans le corpus W. La prédiction d'un corpus de test est d'autant plus précise que la perplexité calculée sur ce corpus est faible. S'agissant d'un modèle stochastique, 2 valeurs de perplexité distinctes, PP et PP^*, peuvent être calculées, selon que $\mathcal{L}(W)$ est pris égal à $\sum_S \mathcal{L}(W,S)$ ou $\mathcal{L}(W,S^*)$. La différence $PP^* - PP$ mesure le degré de stochasticité du modèle, ou encore, dans quelle mesure la précision de la prédiction se dégrade lorsque la vraisemblance de la phrase est approximée avec la vraisemblance de la meilleure segmentation (comme c'est le cas dans les systèmes de reconnaissance). Les expériences de la section 3.2 consistent à évaluer la différence $PP^* - PP$, puis à étudier l'influence de la procédure d'estimation, (3) ou (4), en termes de perplexité et de taille du modèle (nombre de 2-uplets de séquences distincts conservés). Puis, les performances sont comparées avec celles de modèles n-grammes (la taille du modèle est alors le nombre de n-uplets de mots distincts observés).

Protocole d'apprentissage Les probabilités bi-multigrammes sont initialisées d'après les fréquences relatives de tous les 2-uplets de séquences de 1, 2, ... n mots observés dans le corpus, et elles sont réestimées au cours de 6 itérations. Pour réduire le risque de sur-apprentissage, les séquences dont le nombre d'occurences initial, et réestimé, est inférieur, respectivement, à 20 et à 10, sont retirées de l'inventaire des séquences possibles - à l'exception des séquences de longueur 1, dont le nombre d'occurences est mis égal à 1. Les probabilités bi-multigrammes et n-grammes [2] sont lissées par backoff [7] avec la méthode de Witten-Bell [9].

Corpus Les expériences sont réalisées sur le corpus anglais "Travel Arrangement" d'ATR (cf. Tab. 1), consistant en dialogues semi-spontanés au sujet d'informations touristiques. Toutes les hésitations ont été remplacées par un marqueur commun "*uh*".

	apprentissage	test
Nb phrases	13 650	2 430
Nb mots	167 000	29 000 (1 % OOV)
Vocabulaire	3 525	+ 280 OOV

Table 1: Corpus "Travel Arrangement" d'ATR

3.2. *Expériences sans classification*

Degré de stochasticité (Tab. 2): la différence $PP^* - PP$ (après apprentissage forward-backward) reste en-deçà d'environ un point de perplexité: approximer la vraisemblance du corpus de test avec

[1] Par exemple, les séquences "thank you for" et "thank you very much for" ne peuvent pas appartenir à une même classe, puisqu'étant de longueurs différentes elles correspondent à des successions de classes de mots distinctes.

[2] Les probabilités n-grammes sont estimées avec le toolkit CMU [3]

celle de la meilleure segmentation ne devrait donc pas détériorer significativement la précision de la prédiction.

n	1	2	3	4
PP	56.0	43.9	44.2	45.0
PP^*	56.0	45.1	45.4	46.3

Table 2: Degré de stochasticité.

Procédure d'estimation (Tab. 3): le choix de la procédure d'estimation influence peu les valeurs de perplexité, la procédure forward-backward procurant un bénéfice très limité. En revanche, la taille du modèle est inférieure d'environ 30% avec l'apprentissage forward-backward: environ 40 000 bi-multigrammes versus 60 000, pour une même valeur de perplexité. Les résultats indiquent par ailleurs que l'heuristique d'abandon des séquences ne permet pas d'éviter totalement le sur-apprentissage, puisque les valeurs de perplexité avec n =3, 4, sont en légère hausse [3].

Perplexité de test PP^*				
n	1	2	3	4
F.-B.	56.0	45.1	45.4	46.3
Viterbi	56.0	45.7	45.9	46.2
Taille du modèle				
n	1	2	3	4
F.-B.	32505	44382	43672	43186
Viterbi	32505	65141	67258	67295

Table 3: Procédure d'estimation forward-backward (F.-B.) ou Viterbi.

Comparaison avec des n-grammes (Tab. 4) La perplexité bi-multigramme la plus faible (43.9) est plus élevée que la perplexité 3-gramme (40.4), mais représente une nette amélioration par rapport à la perplexité 2-gramme (56.0). Le nombre de 2-uplets de séquences conservés est largement inférieur au nombre de 3-grammes (45000 versus 75000), ce qui illustre la capacité de sélection du modèle.

Perplexité de test PP				
n (and n)	1	2	3	4
n-gr.	314.2	56.0	40.4	39.8
bi-multigr.	56.0	43.9	44.2	45.0
Taille du modèle				
n et n	1	2	3	4
n-gr.	3526	32505	75511	112148
bi-multigr.	32505	44382	43672	43186

Table 4: Comparaison avec des n-grammes

3.3. Expériences avec classification

Les valeurs de perplexité obtenues avec le modèle de classes de bi-multigrammes sont systématiquement supérieures à celles du modèle sans classes. La hausse de perplexité reste en-deçà de 1 point quand une faible fraction des séquences (10 à 20%) sont regroupées en classes, mais se détériore considérablement à mesure que le nombre de classes de la partition est plus fortement restreint. La moindre précision de la prédiction, qui résulte de la classification, est un effet usuellement observé avec les modèles n-grammes, en l'absence d'interpolation entre les modèles avec classes et sans classes. Des expériences avec interpolation sont donc à mener pour permettre de conclure quant au bénéfice potentiel d'une classification des bi-multigrammes. Un avantage attendu de la classification de séquences ayant des longueurs différentes est de réduire l'impact de certaines irrégularités linguistiques caractérisant la parole spontanée, comme les interjections et hésitations. Pour illustrer ce point, les séquences regroupées les premières dans une même classe pendant la classification initiale sont présentées dans le tableau 5. Les séquences différant par l'insertion d'une hésitation (*uh*) sont fréquemment regroupées dans une même classe (cf. le haut du tableau). La partie basse du tableau 5 illustre par ailleurs la cohérence sémantique des classes obtenues sans supervision. Cette cohérence incite à explorer les possibilités de cette approche pour des applications en compréhension du langage, des informations de nature sémantique/pragmatique (étiquetage en actes de langage par exemple) pouvant dans ce cas contraindre le processus de classification.

[3] n =4 signifie que les dépendances prises en compte peuvent s'étendre sur 8 mots

{ yes_that_will_be ; *uh*_yes_that's }
{ yes_*uh*_i ; i_see_i }
{ okay_i_understand ; *uh*_yes_please }
{ could_you_recommend ; *uh*_is_there }
{ *uh*_could_you_tell ; and_could_you_tell }
{ so_that_will ; yes_that_will ; yes_that_would ; uh*_that_would }
{ if_possible_i'd_like ; we_would_like ; *uh*_i_want }
{ that_sounds_good ; *uh*_i_understand }
{ *uh*_i_really ; *uh*_i_don't }
{ *uh*_i'm_staying ; and_i'm_staying }
{ all_right_we ; *uh*_yes_i }

{ good_morning_this ; good_afternoon_this }
{ yes_i_do ; yes_thank_you }
{ we'll_be_looking_forward ; we_look_forward }
{ for_your_help ; for_your_information }
{ hold_the_line ; wait_for_a_moment }
{ yes_that_will_be ; and_could_you_tell }
{ please_go_ahead ; you_like_to_know }
{ what_time_would_you ; and_you_would }
{ yes_there_is ; but_there_is }
{ john_phillips_in_room ; ms._suzuki_in }
{ name_is_suzuki ; name_is_ms._suzuki }
{ name_is_mary_phillips ; name_is_kazuo_suzuki }
{ hold_on_a_moment ; wait_a_moment }
{ give_me_some ; also_tell_me }

Table 5: Exemple de séquences regroupées dans une classe commune "{ ... ; ... ; ... }"

$$p^{(k+1)}(s_j \mid s_i) = \frac{\sum_{t=1}^{T} \alpha(t,\ l_i)\ p^{(k)}(s_j \mid s_i)\ \beta(t+l_j,\ l_j)\ \delta_i(t-l_i+1)\ \delta_j(t+1)}{\sum_t \alpha(t,\ l_i)\ \beta(t,\ l_i)\ \delta_i(t-l_i+1)}$$

l_i est la longueur de la séquence s_i, et la variable de Kronecker $\delta_k(t)$ vaut 1 si la séquence commençant avec le mot de rang t est s_k, et 0 sinon.

Table 6: Formule de réestimation forward-backward bi-multigramme

4. Conclusion

Un algorithme permettant d'estimer la distribution de séquences de longueurs variables entretenant des dépendances bigrammes, et permettant optionnellement de classer ces séquences, a été proposé pour la modélisation statistique du langage. Des expériences sur un corpus de dialogues semi-spontanés d'informations touristiques montrent que les valeurs de perplexité obtenues restent supérieures à la perplexité 3-gramme, mais avec une réduction significative de la taille du modèle. La procédure d'estimation devrait pouvoir être améliorée par la définition d'heuristiques permettant de mieux réduire la tendance au sur-apprentissage. Pour accroître également la capacité de généralisation du modèle, des expériences d'interpolation entre des modèles avec et sans classification sont à mener. Enfin, la pertinence sémantique des classes de séquences, incite à explorer les possibilités offertes par cette approche pour des applications en compréhension du langage.

Annexe: Algorithme Forward-backward

Dans cette annexe, nous présentons l'algorithme forward-backward utilisé au cours de nos expériences pour implémenter la formule (3), dans le cas $\overline{n} = 2$. L'algorithme repose sur la définition d'une variable forward:

$$\alpha(t, l_i) = \mathcal{L}(W_{(1)}^{(t-l_i)}\ [W_{(t-l_i+1)}^{(t)}])$$

représentant la vraisemblance des t premiers mots du corpus, les l_i derniers mots formant une séquence,

et d'une variable backward:

$$\beta(t, l_j) = \mathcal{L}(W_{(t+1)}^{(T)} \mid [W_{(t-l_j+1)}^{(t)}])$$

représentant la vraisemblance conditionnelle des $(T-t)$ derniers mots, sachant qu'ils sont précédés de la séquence $[w_{(t-l_j+1)} \ldots w_{(t)}]$. En supposant la vraisemblance d'une segmentation calculée selon (2), l'équation (3) se réécrit comme indiqué dans la Tab. 6. Les variables α et β sont calculées par récurrence:

pour $1 \leq t \leq T+1$:

$$\alpha(t, l_i) = \sum_{l=1}^{n} \alpha(t-l_i, l)\ p([W_{(t-l_i+1)}^{(t)}] \mid [W_{t-l_i-l+1}^{(t-l_i)}])$$

$\alpha(0,1) = 1,\ \alpha(0,2) = \ldots = \alpha(0,n) = 0.$

pour $0 \leq t \leq T$:

$$\beta(t, l_j) = \sum_{l=1}^{n} p([W_{(t+1)}^{(t+l)}] \mid [W_{(t-l_j+1)}^{(t)}])\ \beta(t+l, l)$$

$\beta(T+1,1) = 1,\ \beta(T+1,2) = \ldots = \beta(T+1,n) = 0.$

Pour une vraisemblance calculée selon (5), le terme $p^{(k)}(s_j \mid s_i)$ dans la formule de réestimation est à remplacer par $p^{(k)}(C_{q(s_j)} \mid C_{q(s_i)})\ p^{(k)}(s_j \mid C_{q(s_j)})$. De même, dans la formule de récurrence de α, le terme $p([W_{(t-l_i+1)}^{(t)}] \mid [W_{t-l_i-l+1}^{(t-l_i)}])$ est à remplacer par la probabilité bigramme de classes correspondante, multipliée par la probabilité conditionnelle de $[W_{(t-l_i+1)}^{(t)}]$ sachant sa classe. Un changement similaire affecte la formule de récurrence de β.

Bibliographie

[1] F. Bimbot, R. Pieraccini, E. Levin, and B. Atal. Modèles de séquences à horizon variable: Multigrammes. In *Proceedings of the XX^{th} JEP*, Trégastel France, June 1994.

[2] P.F. Brown, V.J. Della Pietra, P.V. de Souza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[3] P. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. *Proceedings of EUROSPEECH 97*, 1997.

[4] S. Deligne and F. Bimbot. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. *Proceedings of ICASSP 95*, 1995.

[5] S. Deligne, F. Yvon, and F. Bimbot. Introducing statistical dependencies and structural constraints in variable-length sequence models. In *Grammatical Inference : Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence 1147, pages 156–167. Springer, 1996.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society*, 39(1):1–38, 1977.

[7] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. on Acoustic, Speech, and Signal Processing*, 35(3):400–401, March 1987.

[8] K. Ries, F. D. Buo, and A. Waibel. Class phrase models for language modeling. *Proceedings of ICSLP 96*, 1996.

[9] I.H. Witten and T.C. Bell. The zero-frequency problem: estimating the probabilities of novel events in adaptative text compression. *IEEE Trans. on Information Theory*, 37(4):1085–1094, July 1991.

L'analyse syntaxique robuste de dialogues transcrits : de la parole aux actes en passant par l'écrit

SYLVAIN DELISLE [1], NARJES BOUFADEN [2], BERNARD MOULIN [2,3]

(1) Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Québec, Canada
e-mail : Sylvain_Delisle@uqtr.uquebec.ca
(2) Département d'informatique, Université Laval, Québec, Canada
(3) Centre de recherche en géomatique de l'Université Laval
e-mail : {boufaden, moulin}@ift.ulaval.ca

ABSTRACT

We present some key elements of a robust syntactic parser designed to handle well- and ill-formed inputs. These inputs are French spoken utterances from task-oriented dialogues that have been independently transcribed. The salient features of our work are: a judicious use of surface clues; the modelling of spoken utterances as distorted written equivalents; and the implementation of a robust parser based on core and peripheral grammar layers which are both controlled by a supervising module. The grammar has been developed in the context of Chomsky's GB Theory and is implemented in Prolog as a (quasi-)deterministic bottom-up RL(k) parser. This work is part of an ongoing research project whose goal is to base the construction of a discourse model on the detection and interpretation of its surface clues.

1. INTRODUCTION

Les travaux que nous rapportons dans cet article portent sur l'analyse syntaxique d'énoncés transcrits de la langue française parlée. Ces énoncés proviennent de différents dialogues réels orientés tâche qui ont été cumulés dans des corpora qui nous servent de matériau de base. Un des corpora qui nous a le plus intéressés récemment est celui de conception de figures géométriques constitué par Ozkan [Ozk94] et analysé par Colineau [Col94]. Ce corpus met en scène un instructeur et un manipulateur, qui n'étaient pas dans la même pièce, mais communiquaient entre eux grâce à des microphones et partageaient les mêmes informations sur leurs écrans (les actions exécutées par l'un étaient visibles par l'autre). L'expérience s'est déroulée dans le contexte de l'utilisation d'un logiciel de dessin. L'instructeur avait sous les yeux la scène à représenter et donnait des directives au manipulateur qui ignorait ce qu'il devait dessiner. Les dialogues ont été enregistrés et filmés et leurs transcriptions orthographiques ont été faites par Ozkan. Voici un extrait de ce corpus (i=instructeur, m=manipulateur) :

(i) alors tu prends euh une petite barre verticale
(i) tu mets à gauche du du rond
(i) tu le mets euh
(m) *geste de déplacement*
(i) un peu plus avant un peu plus avant aussi
(m) un peu plus bas ouais.

Le projet MAREDI (MArqueurs et REprésentation des DIscours), dans lequel s'inscrivent nos travaux, vise à développer une approche et un outil d'analyse de discours qui s'appuie essentiellement sur la détection et l'interprétation de marques de surface afin d'élaborer un modèle conceptuel du discours sous la forme d'un ensemble d'états mentaux structurés ([Mou97]). Pour construire automatiquement le modèle conceptuel du discours à partir de l'analyse des énoncés contenus dans un dialogue, nous développons actuellement un système constitué de quatre composantes principales : un réseau neuronal qui sert à déterminer les types d'actes de dialogue accomplis à partir d'indices syntaxiques relevés à la surface des énoncés ([Col96]; [Col97]); un analyseur syntaxique robuste qui doit pouvoir bien réagir aux problèmes d'extragrammaticalité qui sont fréquents dans l'oral, et un analyseur sémantique mettant en œuvre une approche à base de cas sémantiques et ayant recours à l'utilisation de templates ([Bou97]); et, finalement, un intégrateur qui, à partir des structures produites par les trois autres composantes, procède graduellement à la construction du modèle conceptuel du discours analysé. Dans ce modèle, les actes de discours (ou de langage) jouent un rôle important : d'où la partie du titre de cet article qui dit "de la parole aux actes". Nos travaux n'incluent pas le traitement du signal sonore et la reconnaissance de la parole, ni leur possible intégration au processus d'analyse subséquent—à cet effet, voir, entre autres, [Gia88]; [Bag92]; [War94]; la section 1.8 de [Col95]; et [Sto97]. Notre recherche est axée sur le traitement automatique post-reconnaissance et, plus spécifiquement, sur l'analyse robuste d'énoncés oraux transcrits. Nous croyons néanmoins que de tels travaux sont d'un grand intérêt pour les chercheurs en traitement de la parole. En effet, nous remettons en cause deux idées généralement bien acceptées dans les recherches reliées au traitement automatique de la langue parlée, soit : 1) que les marques de surface sont trop peu fiables pour constituer la base d'un modèle d'analyse; et 2) que la langue écrite est trop différente de la langue parlée pour que les grammaires et analyseurs développés pour la première soient vraiment utiles au traitement de la seconde. Dans ce qui suit, le deuxième point constitue le thème principal de nos propos.

2. ANALYSE SYNTAXIQUE DE L'ORAL

L'analyse syntaxique robuste ([Kie96]; [Van95]) intéresse autant les chercheurs travaillant sur l'écrit (par ex., [Hob92]; [Lin93]) que ceux travaillant sur l'oral (par ex., [Gia88]; [Lav96]). Cependant, ceux qui travaillent sur l'analyse de l'oral, surtout l'oral spontané et ses nombreux phénomènes caractéristiques—répétitions, par ex.[Lan90]; réparations, par ex. [Hee97]; hésitations, par ex. [Nak94]; etc.—sont confrontés à une difficile question : dans une perspective de traitement automatique, est-il préférable de considérer l'oral comme une langue autonome ou comme une altération de l'écrit ? La première alternative nous oblige à créer des ressources (lexiques, grammaires et analyseurs) spécifiques à l'oral et difficilement transportables à différents domaines d'application ou adaptables au traitement de l'écrit. La seconde alternative nous questionne sur la façon de modéliser et traiter l'écrit "altéré" (l'oral) afin de préserver le meilleur des deux mondes dans l'ensemble de nos

ressources initialement conçues pour l'écrit. Nous croyons que la seconde alternative est en général la plus intéressante et qu'elle mérite peut-être plus d'attention qu'elle n'en a reçue jusqu'ici. D'autant plus qu'historiquement, de nombreux travaux et ressources informatiques ont été développés d'abord pour la langue naturelle écrite et ne demandent qu'à être mieux utilisés aux fins du traitement de l'oral. Notre approche à l'analyse syntaxique robuste de l'oral se définit sur la base d'un noyau grammatical conçu pour l'écrit. Précisons qu'il s'agit d'une approche linguistique, basée sur une représentation des connaissances en termes de règles explicites, par opposition à une approche statistique (par ex. [Mag94]).

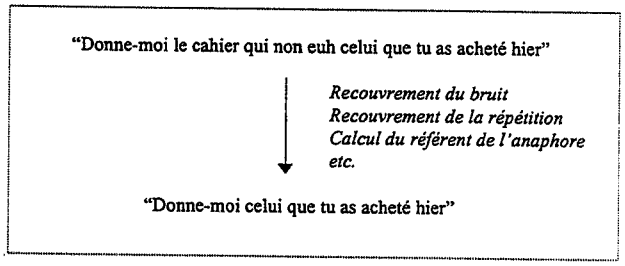
2.1 Écrit et oral : noyau et périphérie

Nous avons basé la conception de notre analyseur syntaxique robuste sur la distinction entre le noyau d'une grammaire et sa périphérie ([Cha93]). Le noyau est chargé de la couverture des phénomènes syntaxiques fondamentaux, généralement associés aux phrases écrites bien formées, tandis que la périphérie est chargée des divers phénomènes qui se démarquent de ceux couverts par le noyau et qui sont souvent associés à la langue parlée. Nous obtenons donc un analyseur en deux couches qui sont gérées par le superviseur dont il est question à la section 2.2.

Pour la définition du noyau de la grammaire, nous avons d'abord construit un ensemble de règles de grammaire couvrant les constructions correctes de nos corpora, c'est-à-dire les constructions bien formées par rapport à une grammaire standard du français écrit. La description de ces règles s'est effectuée à l'aide du formalisme de la Théorie X-barre de Chomsky—la modélisation de la grammaire s'est faite dans le contexte de la Théorie du Gouvernement et du Liage ([Cho81]). L'utilisation de nos corpora nous a aussi permis de valider les règles de grammaire et de les ordonner selon leur probabilité relative d'application. En plus de ces règles de grammaire de base, l'analyse syntaxique requiert des règles de transformation qui permettent de reconnaître différentes constructions telles que les constructions interrogatives, relatives, complétives, etc. Le déclenchement des transformations s'inspire de la méthode utilisée pour la détermination des actes de discours de Colineau [Col96], soit par le biais de la détection de marques lexicales ou syntaxiques. L'ensemble de ces deux types de règles constitue donc l'analyseur syntaxique noyau. Au niveau de sa réalisation informatique, il est implémenté en Prolog et utilise une stratégie (quasi-)déterministe ascendante de type RL(k). Le parcours de la droite vers la gauche facilite le traitement des distorsions de l'oral et permet d'éviter le *backtracking* en cours d'analyse. Le lexique utilisé par l'analyseur est BDLEX de l'IMAG à Grenoble. C'est un lexique de quelques 20Mo pour le français standard. La couche périphérique de l'analyseur syntaxique est définie à partir d'un ensemble d'heuristiques et de mécanismes de recouvrement qui interviennent à chaque fois que le noyau de l'analyseur syntaxique est dans l'impossibilité de produire une analyse. Ces techniques de recouvrement modifient l'état original de l'énoncé afin de favoriser la progression de son analyse. Cette stratégie est cohérente avec la perspective selon laquelle l'oral est vu comme une altération ou une distorsion de l'écrit. La gestion des techniques de recouvrement est effectuée grâce à un module de supervision (ou superviseur) qui incarne la robustesse de l'analyseur syntaxique.

2.2 Le superviseur : la robustesse

Le superviseur détecte les distorsions et il ordonnance les opérations nécessaires à l'analyse syntaxique et au recouvrement des distorsions : en fait, en simplifiant un peu, il tente de trouver l'énoncé bien formé qui est à l'origine de l'énoncé oral comportant des distorsions. Les formes les plus fréquentes de distorsion dans nos corpora sont les suivantes : les bruits (constituant en trop), les répétitions, les effacements (absence d'un constituant), les interruptions et les ellipses[1]. À titre d'exemple, voici la formulation de l'heuristique principale qui sert à traiter le cas des répétitions : "seul le dernier syntagme, partie de syntagme, ou unité lexicale de la répétition est retenu et utilisé pour la reprise de l'analyse syntaxique". L'exemple suivant illustre comment un énoncé comportant plusieurs distorsions est modifié par la couche périphérique pour être ensuite passé à la couche noyau et être analysé avec succès.



Dans le but de décrire plus précisément le fonctionnement du superviseur, voici toutes les opérations contrôlées par le superviseur qui permettent l'analyse syntaxique réussie de l'énoncé *donne-moi le cahier qui non euh celui que tu as acheté hier* (nous rappelons que le parcours se fait de la droite vers la gauche) :

1- L'absence du syntagme nominal objet du verbe *acheter* cause la mise en attente des transformations inverses clitique, relative et de la procédure de calcul d'ellipse.

2- La présence d'un pronom relatif que cause l'exécution de la transformation inverse de la transformation relative puisque c'est la seule opération qui a un ensemble complet de marques. Si la marque disparaît après la transformation, alors il y a exclusion des deux autres transformations inverses clitique et la procédure de calcul d'ellipse.

3- La détection du pronom *celui* engendre la mise en attente de la transformation inverse de la pronominalisation.

4- La détection d'un adverbe, constituant ne pouvant être inséré dans le syntagme courant, cause le déclenchement de la procédure de recouvrement de bruit pour les mêmes raisons que celles de l'étape 2.

5- La détection d'un autre complémenteur *qui,* marque identifiant la transformation relative, cause la mise en attente de la transformation correspondante. De plus, la procédure de recouvrement de bruit est aussi activée puisque ce constituant ne respecte pas la condition d'insertion[2]. Cependant, le déclenchement du recouvrement du bruit élimine la marque, donc toutes les trans-

[1] Les ellipses ne sont pas à proprement parlé des distorsions. Nous les associons aux distorsions pour simplifier la présentation.

[2] Un constituant respecte la condition d'insertion si l'une de ces trois conditions est vérifiée : il peut être inséré dans le syntagme en cours de construction ; il peut représenter le début d'un nouveau syntagme ; il représente à lui seul un syntagme.

formations mises en attente par cette marque sont ignorées.

6- La détection d'un syntagme nominal *le cahier,* ainsi que la marque détectée à l'étape 3, engendre la résolution d'une anaphore. Mais la présence de deux syntagmes nominaux se chevauchant engendrent aussi une procédure de recouvrement de la répétition. Les deux opérations pouvant s'exécuter simultanément engendre un conflit d'exécution qui est résolu par l'ordre d'occurrence des marques. En effet, c'est l'anaphore qui sera déclenchée en premier lieu puisque celle-ci avait déjà une marque détectée dans les lectures précédentes. Ainsi, par ordre de précédence, c'est la procédure de calcul d'anaphore qui est déclenchée en premier. Ensuite, la marque étant encore présente après le calcul de l'anaphore, la procédure de recouvrement de la répétition est exécutée, ce qui termine correctement l'analyse syntaxique.

L'algorithme qui contrôle le superviseur permet de résoudre les problèmes de précédence de manière assez simple. De plus, l'ajout de nouvelles opérations pour la gestion d'autres distorsions n'entraîne pas la détermination d'un nouvel ordre de précédence car celui-ci s'effectue de manière dynamique à partir des marques présentes dans l'énoncé courant. Le superviseur utilise un module subalterne, appelé contrôleur des marques, qui se charge de détecter les marques qui permettent d'activer les mécanismes de recouvrement. Le contrôleur de marques utilise le concept de point de contrôle, c'est-à-dire une étape précise du processus d'analyse à laquelle il vérifie la progression de l'analyse et teste la présence des marques. Les trois points de contrôle que nous avons définis sont : à la lecture d'une nouvelle unité lexicale *(Shift)*; à la formation d'un syntagme verbal; et à la fin de l'analyse.

3. PREMIERS RÉSULTATS

L'implémentation actuelle de notre analyseur syntaxique robuste ne couvre pas tous les phénomènes linguistiques, ni de l'écrit, ni de l'oral. La première phase de ces travaux visait d'ailleurs à démontrer la valeur potentielle de l'approche; nous estimons avoir atteint cet objectif. Les transformations du noyau qui ont été implémentées et testées sont celles des pronoms clitiques, des interrogatives, des négatives et des relatives. Une partie seulement du traitement de toutes les distorsions possibles est actuellement en fonction : implémentation complète pour les bruits et répétitions, partielle pour les effacements, et rien encore pour les ellipses et interruptions. Nous avons récemment évalué cette implémentation sur deux corpora, d'une centaine d'énoncés chacun, issus de la mise en situation de dessin à laquelle nous faisions allusion plus haut. Il faut bien entendu interpréter ces premiers résultats (voir Tableau 1) avec prudence : la relative disparité des résultats témoigne des différences dans la nature des énoncés composant chacun des corpora. Dans l'ensemble, nous les trouvons satisfaisants pour un premier prototype, notamment les résultats pour le bruit qui est un phénomène présent dans plus de la moitié des énoncés de chaque corpus. Cependant, les résultats pour "zéro distorsions" (énoncés corrects) nous ont étonnés : ils étaient plus faibles qu'attendus. Après examen plus détaillé de leurs causes, nous avons constaté qu'ils étaient causés principalement, soit par une erreur d'implémentation, soit par des lacunes au niveau des phénomènes linguistiques couverts par le noyau de l'analyseur (par ex. certaines constructions interrogatives). Ce dernier constat indique peut-être que nos efforts se sont trop concentrés sur le traitement des distorsions, au point de perdre de vue l'importance du rôle du noyau de l'analyseur. D'autant plus que le noyau joue un rôle déterminant dans l'évaluation des capacités de l'analyseur pour le traitement de l'écrit et, donc, dans l'évaluation finale de notre hypothèse de travail portant sur l'oral par rapport à l'écrit. Ces problèmes seront corrigés dans les mois à venir et nous sommes confiants de la qualité des résultats d'analyse que nous pourrons alors obtenir.

Phénomène	Corpus C11	Corpus C5
zéro distorsion	63%	33%
bruit	100%	75%
effacement	100%	50%
répétition	86%	n/a

Tableau 1 : Taux de succès du traitement des énoncés comportant une occurrence de chacun des phénomènes considérés par l'analyseur syntaxique robuste. 'n/a' signifie qu'il n'y avait pas d'énoncé de ce type.

4. CONCLUSION ET TRAVAUX FUTURS

Nous avons présenté les principales caractéristiques de notre approche à l'analyse syntaxique robuste d'énoncés transcrits de l'oral. Les caractéristiques originales en sont l'exploitation judicieuse des marques de surface; l'adoption d'une perspective selon laquelle la langue orale est une variation de la langue écrite; et la réalisation d'un analyseur syntaxique robuste à partir des concepts de grammaire noyau et périphérique, et d'un superviseur qui contrôle ces deux couches. Nous signalons au lecteur deux recherches en rapport direct avec nos propres travaux. Premièrement, Langer ([Lan90]) qui s'intéresse spécifiquement à l'analyse syntaxique d'énoncés oraux dits "mal-formés" et utilise un concept de normalisation qu'il définit comme étant la relation entre un énoncé malformé et sa version bien formée. Pour ce faire, il augmente un analyseur de l'écrit d'un certain nombre d'heuristiques qui permettent de traiter quelques phénomènes comme les déviations morpho-syntaxiques, certaines réparations et certaines répétitions. Et deuxièmement, Nakano *et al.* ([Nak94]) qui augmentent un analyseur d'abord conçu pour l'écrit d'un ensemble de règles qui permettent d'élargir la couverture des phénomènes syntaxiques en ajoutant ceux qui sont spécifiques à la langue orale Japonaise. Par rapport à ces travaux, notre approche présente le net avantage d'établir une distinction claire, à la fois du point de vue théorique et du point de vue de la réalisation informatique, entre ce que nous désignons comme étant le noyau et la périphérie de la grammaire. Qui plus est, le contrôle de la coordination de ces deux couches est effectué par un module dédié à cette tâche, le superviseur : il s'agit donc d'une approche plus raffinée que celle qui consiste simplement à ajouter des règles à une grammaire existante.

L'état d'avancement de notre projet est le suivant : l'analyseur neuronal est terminé, une première version de l'analyseur syntaxique robuste et de l'analyseur sémantique a été réalisée, et le développement de l'intégrateur du modèle du discours débutera sous peu. En ce qui concerne l'analyseur syntaxique, nos prochains travaux seront principalement orientés vers les points suivants : 1) augmenter la couverture de la grammaire noyau; 2) raffiner les heuristiques implantées afin d'éviter certaines ambiguïtés entre les procédures de traitement des distorsions et de la grammaire; 3) augmenter le

nombre de distorsions traitées et ce, en identifiant des heuristiques pour les interruptions; 4) implanter une méthode de calcul des ellipses qui permettrait de traiter plusieurs énoncés problématiques (au moins 20%); et 5) analyser d'autres cas de distorsions présents dans d'autres de nos corpora (par ex. l'inversion des arguments comme dans l'énoncé *en tout tu m'as dit il en faut combien*). À plus long terme, il serait intéressant de permettre à l'analyseur syntaxique d'avoir directement accès à d'autres sources d'information comme certaines données sur la prosodie, les résultats du réseau neuronal et peut-être même de prévoir un protocole de communication avec l'analyseur sémantique.

REMERCIEMENTS

Nous remercions le FCAR (Fonds pour la formation de Chercheurs et l'Aide à la Recherche) ainsi que le CRSNG (Conseil de Recherche en Sciences Naturelles et Génie du Canada) pour l'aide financière apportée à nos travaux de recherche.

BIBLIOGRAPHIE

[Bag92] Baggia P., E.Gerbino, E.Giachin & C.Rullent (1992), "Real-time Linguistic Analysis for Continuous Speech Understanding", *Proc. of the 3rd Conf. on Applied Natural Language Processing,* Trento, Italy, 33-39.

[Bou97] Boufaden N., S.Delisle, B.Moulin & M.Gouiaa (1997), "Linguistique informatique et robustesse : analyse syntaxique et sémantique pour la modélisation de dialogues finalisés", *Actes de la Conf. FRACTAL-97*, Besançon, France, 75-84.

[Cha93] Chanod J.-P. (1993), "Problèmes de robustesse en analyse syntaxique", *Actes de la Conf. Informatique et Langue Naturelle ILN'93*, 223-244.

[Cho81] Chomsky N. (1981), *Lectures on Government and Binding*, Dordrecht : Foris.

[Col94] Colineau N. (1994), "Vers une compréhension des actes de discours", rapport de DEA en Sciences Cognitives, Institut National Polytechnique de Grenoble.

[Col95] Cole R.A., J.Mariani, H.Uszkoreit, A.Zaenen & V.Zue, edts. (1995), *Survey of the State of the Art in Human Language Technology*, http://www.cse.ogi.edu/CSLU/HLTsurvey/, 590p.

[Col96] Colineau N. & B.Moulin (1996), "Un modèle connexionniste pour la reconnaissance d'actes de dialogue", *Actes de la conf. Informatique et Langue Naturelle ILN'96*, Nantes, France, 157-174.

[Col97] Colineau N. (1997), "Étude des marqueurs discursifs dans le dialogue finalisé", thèse de 3e cycle de l'Université Joseph Fourier, Spécialité Sciences Cognitives, décembre 1997, 283p.

[Gia88] Giachin E.P. & C.Rullent (1988), "Robust Parsing of Severely Corrupted Spoken Utterances", *Proc. of the 12th International Conf. on Computational Linguistics (COLING-88),* Budapest, Hungary, Volume 1, 196-201.

[Hee97] Heeman P.A. & J.F.Allen (1997), "Intonational Boundaries, Speech Repairs and Discourse Markers: Modeling Spoken Dialog", *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL,* UNED, Madrid, Spain, 254-261

[Hob92] Hobbs J.R., D.E.Appelt, J.Bear, M.Tyson & D.Magerman (1992), "Robust Processing of Real-World Natural-Language Texts", in (edts.) Paul S. Jacobs, *Text-Based Intelligent Systems : Current Research and Practice in Information Extraction and Retrieval*, LEA, 13-33.

[Kie96] Kiezebrink R. & H.deVreught (1996), "An Annotated Bibliography on Robust Parsing and Related Topics", technical report, Faculty of Technical Mathematics and Informatics, Delft Univ. of Technology (The Netherlands).

[Lan90] Langer H. (1990), "Syntactic Normalization of Spontaneous Speech", *Proc. of the 13th International Conf. on Computational Linguistics (COLING-90),* Helsinki, Finland, Volume 3, 180-183

[Lav96] Lavie A. (1996), "GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language", Ph.D. Thesis, School of Computer Science, Carnegie Mellon University (USA).

[Lin93] Linebarger M.C., L.M.Norton & D.A.Dahl (1993), "A Portable Approach to Last Resort Parsing and Interpretation", *Proc. of the 1993 Human Language Technology Work.*, 31-36.

[Mag94] Magerman D.M. (1994), "Natural Language Parsing as Statistical Pattern Recognition", Ph.D. Thesis, Dept. of Computer Science, Stanford University (USA).

[Mou97] Moulin B.& D.Rousseau (1997), "An Approach for Modelling and Simulating Conversations", à paraître dans D. Vanderveken et S. Kubo (edts.), *Essays in Speech Act Theory*, John Benjamins. Aussi, rapport de recherche, Univ. Laval, mai 1994.

[Nak94] Nakano M., A.Shimazu & K.Kogure (1994), "A Grammar and a Parser for Spontaneous Speech", *Proc. of the 15th International Conf. on Computational Linguistics (COLING-94),* Kyoto, Japan, Volume 2, 1014-1020

[Ozk94] Ozkan N. (1994), "Vers un modèle dynamique du dialogue : analyse de dialogues finalisés dans une perspective communicationnelle", thèse de Doctorat en Sciences Cognitives, Institut National Polytechnique de Grenoble.

[Sto97] Stolcke A. (1997), "Linguistic Knowledge and Empirical Methods in Speech Recognition", *AI Magazine*, Special Issue on Natural Language Processing, 18(4), Winter 1997, 25-31.

[Van95] Van der Ende D. (1995), "Robust Parsing: An Overview", technical report, Dept. of Computer Science, Univ. of Twente (The Netherlands).

[War94] Ward W. & S.Issar (1994), "Recent Improvements in the CMU Spoken Language Understanding System", *Proc. of the Human Language Technology Work,* Plainsboro, New Jersey, USA, 213-216.

Evaluation de modèles de langage à base de trigrammes de classes et de mots, avec le Jeu de Shannon.

Michèle Jardino
LIMSI-CNRS
Université Paris-Sud
BP 133, 91403 Orsay Cedex, France
jardino@limsi.fr

ABSTRACT

Tri-gram class-based language models are evaluated using an adaptation of the Shannon Game and compared with a tri-gram word-based model. These models are trained on a very large corpus. The mapping of the untagged words in non-overlapping classes is automatically realized during an iterative process, according to the local context of the words observed in the training data. We show that a combination of class-based and word-based models improve perplexity. The optimum is independant of the number of classes used in the class-based models, when it ranges from 300 to 2 000 classes.

1. INTRODUCTION

Les modèles de langage probabilistes utilisés dans les systèmes de reconnaissance de la parole ne résolvent pas de manière satisfaisante les nombreuses inflexions de la langue française. Bien que les textes d'apprentissage actuellement disponibles soient de taille respectable, ils ne contiennent pas toutes les chaînes de mots que peut générer le français et il est alors difficile de prédire l'apparition d'un mot connaissant son histoire, même si celle-ci est réduite aux deux mots précédents. Des modèles à base de classes ont été créés pour pallier cet inconvénient.

Nous montrons dans cet article ce qu'apporte un modèle construit à partir de trigrammes de classes par rapport à un modèle construit à partir de trigrammes de mots. Nous les évaluons non seulement globalement, en terme de perplexité, mais également en terme de rang d'observation des mots du vocabulaire, suivant une adaptation du Jeu de Shannon [SHA51], [BIM97], [ELB97].

2. ADAPTATION DU JEU DE SHANNON

Lors du Jeu de Shannon [SHA51], il est demandé à un humain de deviner les lettres d'un texte, en commençant ses prédictions par la première lettre puis la seconde et ainsi de suite dans l'ordre d'écriture. La personne donne sa première réponse pour la première lettre, si elle n'est pas correcte, il en donne une seconde et ainsi de suite.

En s'inspirant de ce jeu, il a été montré que l'on pouvait évaluer un modèle de langage en misant sur des mots à deviner dans un texte [BIM97], [ELB97]. Cette méthode a été développée initialement pour évaluer des modèles de langage en dehors et indépendamment des laboratoires où ils ont été construits. Nous l'avons reprise car elle donne plus de renseignements sur les modèles que la seule évaluation de perplexité.

2.1 Méthode d'évaluation

Etant donné un vocabulaire et un texte dans lequel on a effacé certains mots, on doit associer à chaque mot manquant, une liste des mots du vocabulaire, rangés selon leurs probabilités décroissantes, telles qu'elles sont prédites par le modèle à évaluer. Dans chaque liste, il n'y a qu'un mot solution. La perplexité est alors calculée comme l'inverse de la moyenne géométrique des mises qui ont été attribués aux mots solutions. Un protocole a été mis en place pour définir les critères d'application de cette méthode [JAR98].

2.2 Protocole

Dix mille phrases ont été sélectionnées au hasard, dans un corpus différent des corpus sur lesquels sont appris les modèles. Ces phrases ont éte tronquées, également au hasard. Il s'agit alors de trouver le mot qui doit suivre chaque phrase tronquée, parmi tous les mots du vocabulaire. La prédiction des mots manquants est faite en donnant les listes de mots hypothèses classés dans l'ordre décroissant des probabilités conditionnelles avec les mises correspondantes.

Ayant défini une taille de vocabulaire de 20 000 mots, il a été montré que le nombre d'hypothèses à proposer pour chaque phrase tronquée pouvait être réduit à 5 000 [JAR98]. Les probabilités des autres mots sont distribuées uniformément de telle sorte que la somme des probabilités conditionnelles sur tout le vocabulaire soit égale à 1.

Finalement, la perplexité est calculée à partir des probabilités conditionnelles des 10 000 mots solutions à trouver, qu'ils soient ou non dans les 10 000 fois 5 000 hypothèses.

3. MODELES DE LANGUAGE PROBABILISTES A BASE DE TRIGRAMMES

Nous avons construit à partir d'un même corpus d'apprentissage, un modèle de langage à base de trigrammes de mots et plusieurs modèles de langage à base de trigrammes de classes.

3.1 Corpus d'apprentissage

Ce corpus est constitué d'articles du journal *Le Monde* et de dépêches de *l'AFP*, couvrant la période 1987-1997. Ce corpus prétraité par G, Adda et al [ADD97] contient environ 300 millions de mots.

Le vocabulaire choisi pour cette application, est composé des 20 000 mots les plus fréquents dans le corpus d'apprentissage plus le mot inconnu et les symboles de début et fin de phrase. Le nombre de trigrammes de mots associés est d'environ 50 millions, le nombre de bigrammes de mots 10 millions.

3.2 Modèle de language à base de trigrammes de mots

Ce modèle est construit à partir des 50 millions de trigrammes de mots observés. Les probabilités des évènements non observés sont obtenues par interpolation non linéaire. Une quantité fixe est prélevée sur tous les trigrammes de mots observés et la masse obtenue est redistribuée sur l'ensemble des trigrammes de mots possibles suivant une distribution d'ordre inférieur de type bigramme ou unigramme [NEY94].

3.3 Modèles de language à base de trigrammes de classes

Nous avons construit plusieurs modèles de langage correspondant à des nombres différents de classes. Le classement est entièrement automatique [JAR96]. Il est réalisé sur les mots non étiquetés du corpus d'apprentissage décrit ci-dessus. La répartition des mots dans des classes disjointes est réalisée lors d'un processus que nous rappelons brièvement maintenant.

Type de classement Les mots du vocabulaire sont classés selon leur contexte lors d'un processus itératif qui minimise la perplexité du corpus d'apprentissage. La perplexité est calculée à partir des probabilités d'observer un mot étant donné sa classe et la classe du mot qui le précède.

Initialement les mots sont regroupés dans une seule classe. Le nombre de classes dans lesquelles les mots vont se répartir est fixé à l'avance A chaque itération, un mot choisi au hasard essaie une nouvelle classe également choisie au hasard. Ce nouveau classement est accepté si la perplexité a diminué. A chaque pas on examine tous les bigrammes de mots du texte d'apprentissage qui comprennent le mot, que celui-ci soit en première ou deuxième position. Si NC est le nombre de classes choisi au préalable, à chaque pas sont évalués et éventuellement modifiés les 4*NC bigrammes de classes correspondants.

Après optimisation, on mesure les nombres de bigrammes et de trigrammes de classes non nuls observés dans le texte d'apprentissage. Le tableau 1 montre l'accroissement de ces nombres avec le nombre de classes. Ainsi, il y a 10 fois moins de trigrammes de classes que de trigrammes de mots pour une répartition des mots dans 300 classes, ce qui peut être intéressant si on veut un modèle de langage plus performant aussi bien en temps de calcul qu'en espace mémoire.

Tableau 1. Nombres de trigrammes et bigrammes de classes du texte d'apprentissage selon le nombre de classes

nb classes	nb trigrammes	nb bigrammes
100	590 174	9 622
300	4 988 978	85 191
500	9 040 346	221 028
1 000	17 712 575	738 607
2 000	28 616 986	1 989 831
20 003	52 516 540	8 540 298

Modèles triclasses A partir des classements réalisés précédemment, nous avons construit plusieurs modèles de langage tri-classes avec respectivement 100, 300, 500, 1 000 et 2 000 classes. Ces modèles prédisent un mot connaissant sa classe, son poids dans la classe, et les classes des deux mots qui le précèdent.

Les probabilités des évènements non observés sont obtenues avec la même méthode d'interpolation non linéaire que celle utilisée pour la construction du modèle de langage à base de trigrammes de mots [NEY94]. Une quantité fixe est prélevée sur tous les trigrammes de classes observés et la masse obtenue est redistribuée sur l'ensemble des trigrammes de classes possibles suivant les probabilités des bigrammes et des unigrammes de classes.

4. EVALUATION

On réalise l'adaptation du Jeu de Shannon sur les phrases tronquées d'un texte test, en calculant la perplexité PP_{Sh} sur les mots à trouver et en analysant leurs rangs d'observation.

4.1 Texte test

Le test est extrait du journal *Le Monde Diplomatique*, paru entre 1990 et 1996. Il comprend 10 000 phrases tirées au hasard parmi 222 736 . Ces phrases contiennent environ 400 000 mots.

4.2 Résultats

Tableau récapitulatif Dans ce tableau sont regroupés les résultats des expériences réalisées avec les cinq modèles à base de classes (100, 300, 500, 1 000 et 2 000

classes) et avec le modèle à base de mots, qui correspond à 20 003 classes, en supposant que chaque mot du vocabulaire est dans une classe différente. On y a représenté outre la perplexité PP_{Sh}, le pourcentage des mots observés au premier rang, puis du premier rang au cinquième rang, puis hors des 5 000 premières hypothèses, et finalement le rang moyen d'observation.

Tableau 2. Perplexités et nombres de mots observés selon leur rang, donnés par les différents modèles de langage

nb classes	100	300	500	1000	2000	20003
PP_{Sh}	250	204	181	167	146	117
mots au 1^{er} rang	12%	13%	14%	15%	16%	18%
mots dans 5^{iers} rangs	30%	33%	34%	36%	37%	40%
mots en dehors des 5 000 1^{ers}	2%	1,6%	1,5%	1,5%	1,4%	1,6%
rang moyen	306	279	248	245	215	166

On constate que les performances du modèle à base de trigrammes de mots sont toujours supérieures à celles des modèles à base de trigrammes de classes. La perplexité diminue quand le nombre de classes augmente ainsi que le rang moyen.

Nous avions constaté auparavant [JAR96] qu'avec un corpus d'apprentissage de taille dix fois plus faible, un modèle tri-classes avait de meilleures performances que le modèle trigrammes de mots correspondant, si l'on choisissait une répartition en mille classes. Ceci montre clairement que le choix entre l'utilisation d'un modèle à base de trigrammes de mots et l'utilisation d'un modèle à base de trigrammes de classes dépend de la taille des données d'apprentissage.

Distribution en rang Les figures suivantes montrent le nombre de mots observés à chaque rang pour différents modèles, sur deux échelles différentes.

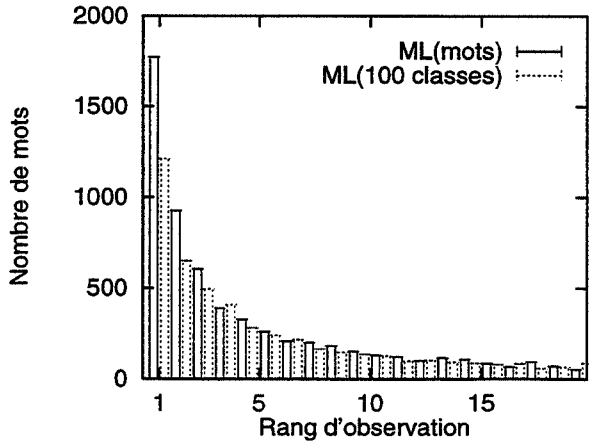


FIG. 1 - *Nombre de mots observés selon leur rang pour un modèle de langage à base de trigrammes de mots, ML(mots) et un modèle de langage à base de trigrammes de classes, ML(100classes)*

La figure 1 est en échelle linéaire. Elle montre que la distribution correspondant au modèle de trigrammes de mots, ML(mots), est la plus pointue. Les distributions des modèles de trigrammes de classes s'étalent davantage. Nous avons choisi de ne représenter sur cette figure que le modèle correspondant à 100 classes, ML(100classes), qui est celui qui s'écarte le plus de la distribution du modèle de trigrammes de mots, ML(mots).

Toutes les distributions en rang des modèles suivent approximativement une loi de Zipf, c'est à dire une distribution inversement proportionnelle au rang d'observation. En exemple, nous avons représenté sur la figure 2, la distribution du modèle ML(100classes) en échelles logarithmiques.

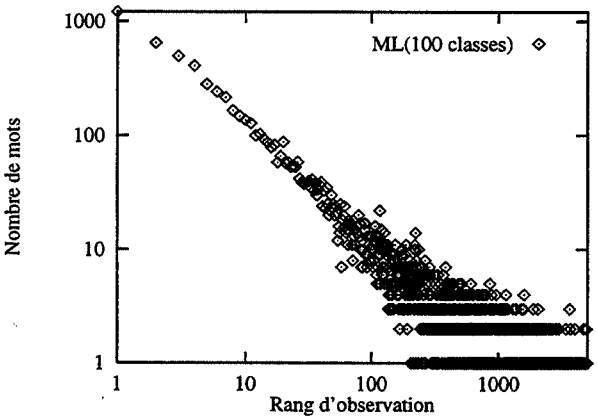


FIG. 2 - *Nombre de mots observés selon leur rang en échelle log-log*

Analyse des différents rangements Dans le tableau 3 nous avons analysé séparément le comportement des mots à trouver, selon que la probabilité conditionnelle donnée par le modèle de trigrammes classes, p_c, est supérieure ou non à celle donnée par le modèle de trigrammes de mots, p_m, ceci pour différents nombres de classes. Nous avons calculé dans chaque cas le rang moyen pour le modèle à base de classes ML(classes), en indiquant entre parenthèses le rang moyen pour le modèle à base de mots ML(mots).

Tableau 3. Nombres de mots respectivement tels que $p_c > p_m$ et $p_c < p_m$ avec les rangs moyens correspondant aux différents ML(classes) étudiés

nb classes	nb mots $p_c > p_m$	rang moyen	nb mots $p_c < p_m$	rang moyen
100	2 978	537 (638)	7 022	343 (72)
300	3 036	470 (608)	6 964	306 (80)
500	**3 060**	436 (587)	6 940	277 (88)
1 000	2 805	427 (601)	7 195	271 (100)
2 000	2 534	409 (618)	7 466	240 (113)

Ce tableau montre qu'il y a un optimum pour le nombre de mots tels que $p_c > p_m$, pour 500 classes. On constate également que le modèle de classes a de meilleures performances essentiellement pour la queue de distribution, puisque le rang moyen est amélioré mais avec des valeurs plus élevées que celles obtenues pour les mots tels que $p_c < p_m$.

Pour illustrer la typologie des classes que nous construisons et ce que peut apporter un modèle construit à partir de ces classes, nous avons étudié, parmi les 140 mots qui sont en dehors des 5 000 premières hypothèses pour ML(mots), le mot *nippons*. Il devait être prédit après les mots *'investissements directs'*. Comme le trigramme *'investissements directs nippons'* n'a pas été observé dans les 300 millions de mots du corpus d'apprentissage, il a une probabilité conditionnelle très faible avec ML(mots) et n'ai pas observé dans les 5 000 premières hypothèses proposées par ce modèle. Par contre le trigramme de classes *'521 287 497'* qui lui est associé par la classification des mots dans 1000 classes, a été observé 68 fois, et *nippons* est prédit par le modèle de classes correspondant, au 234^{me} rang.

L'analyse a posteriori du contenu des classes donne des informations supplémentaires sur cette prédiction. La classe *521* qui contient le mot *investissements*, regroupe des mots du même type, que l'on identifie comme étant majoritairement des mots masculins pluriels. La classe *287* qui contient le mot *directs*, rassemble des adjectifs masculins pluriels. La classe *497* est celle du mot *nippons*, elle contient également des adjectifs masculins pluriels. Ces deux dernières classes ne sont pas regroupées car les adjectifs ont été observés dans des contextes de typologies différentes : la classe *287* étant précédée d'une classe de mots masculins singuliers et suivie d'une classe d'adjectifs masculins singuliers, est différente de la classe *497* qui est précédée d'une classe d'adjectifs masculins pluriels.

4.3 Combinaison des modèles

Pour prendre en compte les avantages de chaque type de modèle, nous avons réalisé la combinaison linéaire des modèles à base de trigrammes de classes avec le modèle à base de trigrammes de mots. La figure 3 donne la perplexité PP_{Sh}, calculée sur les phrases tronquées en fonction du coefficient d'interpolation.

On constate que l'on obtient une perplexité plus basse en combinant à hauteur de 20% un modèle trigramme de classes avec un modèle trigramme de mots. Le point intéressant est que l'optimum est quasiment indépendant du nombre de classes quand ce nombre varie entre 300 et 2 000 classes.

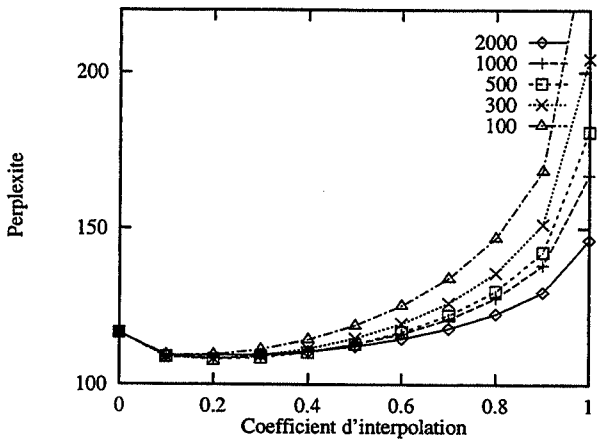


FIG. 3 - *Perplexité du texte test, PP_{Sh}, en fonction du coefficient d'interpolation, pour les modèles à base de trigrammes de classes combinés avec le modèle à base de trigrammes de mots et pour différents nombres de classes (2 000, 1 000, 500, 300, 100)*

5. CONCLUSION

Nous avons montré à l'aide d'une adaptation du Jeu de Shannon qu'un modèle de language de trigrammes de classes améliore les performances d'un modèle de langage de trigrammes de mots, pour un nombre de classes relativement réduit. Etant donné les ambiguités de la langue française qu'un tel modèle peut résoudre, il devrait permettre d'obtenir de meilleures performances en reconnaissance de la parole en français.

REFERENCES

[ADD97] Adda G., Adda-Decker M., Gauvain J.L., Lamel L., "Text Normalization and Speech Recognition in French", EUROSPEECH'97, Rhodes, Greece, p.1711, 1997

[BIM97] Bimbot F., El-Bèze M., Jardino M., "An alternative scheme for perplexity estimation", ICASSP 1997, Munich, Germany.

[ELB97] El-Bèze M. Jardino M., Bimbot F., "Une approche alternative pour le calcul de la perplexité", JST97, Avignon, France.

[JAR96] Jardino M., "Multilingual stochastic n-gram class language models", ICASSP 1996, Atlanta, USA.

[JAR98] Jardino M., Bimbot F., Igounet S., Smaili K., Zitouni I., El-Bèze M., "A first evaluation campaign for language models", LREC 1998, Grenade, Espagne.

[NEY94] Ney, H., Essen U., Kneser, R., "On structuring probabilistic dependences in stochastic language modelling", Computer Speech and Language, 1994 , 8

[SHA51] C. E. Shannon, "Prediction and entropy of printed English", Bell Syst. Techn. J., pp. 50-64, Jan. 1951.

Reconnaissance I

Elements pour la mise au point de système de reconnaissance grand vocabulaire en français

Martine Adda-Decker, Gilles Adda, Jean-Luc Gauvain, Lori Lamel

Groupe Traitement du langage parlé
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{madda,gadda,gauvain,lamel}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In previous studies concerning speech recognition in French, lexical coverage has been identified as a major contributor to recognition error rates. High lexical variety also implies poorer language model (LM) training for a fixed text corpus, even though the LM component is of prime importance in French to discriminate homophones. The use of combined word and class N-gram language models is investigated. To address the coverage problem the system vocabulary has been extended from 20k to 65k words and different text normalisation and selection methods are addressed. Comparing 20k and 65k systems, a relative error reduction of 40% is observed. Word error rates analysed as a function of word frequencies highlight the importance of the LM in the recognition process. Part of this work has been carried out within the ARC *Linguistics, Computer Sciences & Spoken Corpora* supported by the AUPELF-UREF. During the official AUPELF'97 evaluation of French recognizers our 65k system obtained a 11.2% word error rate.

1. INTRODUCTION

Le français est une langue inflectionnelle; cette spécificité, qu'elle partage avec d'autres langues (mais pas avec l'anglais), conduit à une grande variété lexicale, ce qui n'est pas sans causer quelques soucis aux systèmes de reconnaissance de la parole : un grand nombre de mots hors vocabulaires (MHV) et des modèles de langage (ML) avec des perplexités élevées. Et il se superpose à cette variété lexicale un fort taux de mots homophones, qui ne peuvent être discriminés que par le modèle de langage. Une étude comparative portant sur des textes de journaux en français et en anglais a montré que si 20% des mots du texte anglais avaient une transcription phonémique (supposée correcte) ambiguë, ce nombre montait à 75% pour le texte français[Gau94]; et pour illustrer la faiblesse de la couverture lexicale en français, la même étude a montré qu'il fallait doubler la taille du lexique en français pour obtenir un couverture comparable à celle de l'anglais. On peut ainsi mieux mesurer les problèmes que rencontre la reconnaissance en français.

Dans cet article, nous présentons les améliorations que nous avons apportées au système de reconnaissance du LIMSI, afin de lui permettre de traiter efficacement la langue française, en insistant sur l'importance de la modélisation lexicale et syntaxique. En premier lieu, nous abordons les moyens de diminuer le nombre de MHV; nous montrons ensuite l'impact sur l'estimation du ML et sur les résultats de reconnaissance, de l'utilisation de textes de 40 à 255 M de mots. Nous donnerons les résultats que nous avons obtenus avec les systèmes 20k et 65k, sur des corpus de tests avec et sans contrôle du taux MHV, en analysant l'impact que ces MHV produisent sur l'efficacité globale du système. Nous présentons une étude des taux d'erreur en fonction de la fréquence des mots erronés, pour mettre en lumière de quelle manière le ML intervient dans le processus de reconnaissance. Cette étude confirme la nécessité de disposer de modèles de langage plus efficaces, les modèles de type n-gramme étant trop limités dans de nombreuses situations, pour la langue française.

2. LE PROJET "DICTÉE VOCALE" DE L'ARC B1 DE L'AUPELF

Une part importante des travaux décrits dans cet article a été produite lors du développement et de l'évaluation à laquelle nous avons participé dans l'ARC B1 de l'AUPELF-UREF[1], portant sur l'évaluation des systèmes de reconnaissance en français. Différentes catégories d'évaluation ont été définies, se différenciant par la taille du lexique (20k, 65k) et par le fait que le taux de MHV était contrôlé ou non. Une évaluation officielle pour la langue française, menée dans le projet LRE-SQALE [Lam95] avait eu lieu, quelques années auparavant, où le lexique était limité à 20k mots et où le taux MHV était contraint à moins de 2% (ce taux, sans contraintes, devant être de 5 à 6%). Pour l'évaluation AUPELF [Dol97], pour le développement et pour l'évaluation officielle, 2 ensembles de test $\mathcal{T}$ de 600 phrases ont été sélectionnés sans contraite sur le taux MHV. On a ensuite extrait de chaque ensemble $\mathcal{T}$, un sous-ensemble $\mathcal{T}'$ de paragraphes totalisant 300 phrases, les paragraphes étant choisis pour obtenir le taux de MHV minimal. Nous présentons ici les résultats obtenus lors du développement sur les 2 ensembles $\mathcal{T}$ et $\mathcal{T}'$.

3. NOUVELLES AVANCÉES POUR LA LANGUE FRANÇAISE

Lors du projet SQALE précédemment évoqué, un des résultats importants qui avaient été mis en évidence, est que le taux de MHV est très différent pour une même taille de vocabulaire d'une langue à l'autre. Etant donné l'impact de ce taux sur l'efficacité de la reconnaissance, il est donc fondamental d'améliorer la couverture lexicale pour les langues inflectionnelles comme l'allemand ou le français.

[1] AUPELF: Association des Universités Partiellement ou Entièrement de Langue Française - UREF: Université des Réseaux d'Expression Française. ARC:Action de Recherche Concertée.

3.1. *Couverture lexicale*

Nous avons étudié l'impact de 3 phénomènes sur la couverture lexicale : la taille du vocabulaire, la définition d'un mot du lexique, et la sélection des mots du lexique. Pour ce faire, nous avons utilisé différents extraits du corpus de texte du journal *Le Monde* (années 1987 à 1995) :
T_0 : années 1987-88 (40M mots)[2], T_0' : années 1994-95 (40M mots), T_1 : années 1987-95 (185M mots), T_2 : years 1991-95 (105M words).

Taille du vocabulaire La table 1, donnant le taux de MHV pour différentes tailles de vocabulaire, montre que la couverture s'accroît toujours de manière sensible en augmentant la taille du vocabulaire. Ainsi, le taux de MHV passe de plus de 6% pour 20k mots à moins de 2% pour 65k mots, sur le corpus $\mathcal{T}$.

Table 1: Taux de MHV sur le corpus de développement $\mathcal{T}$ pour des listes de mots de taille variant de 10k à 65k mots. La liste est choisie en sélectionnant les N mots les plus fréquents du texte T_0.

lexique	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
% MHV	6,38	4,30	3,15	2,36	1,95	1,79

Définition du mot Pour un vocabulaire de taille fixe, la couverture peut grandement varier suivant la définition que l'on adopte de ce qu'est un "mot". Une étude détaillée de l'impact des différentes définitions peut être trouvée dans [Add97a]. Dans la figure 1, le taux de MHV est réduit de 50% lorsque l'on utilise une normalisation significative (N_b, N_c) du texte, en partant d'un texte (N_a) sans traitement particulier. La normalisation N_b est obtenue à partir du texte N_a en séparant les ponctuations ambiguës (- , ', .), en décapitalisant à bon escient les majuscules de début de phrase, et en normalisant les nombres et les sigles; N_c diffère de N_b en ce qu'aucune disctinction minuscule-majuscule n'a été conservée, que les signes diacritiques ont été retirés, et que ponctuations ambiguës ont été systématiquement décomposées. La forme N_b sera la normalisation utilisée dans les expériences suivantes.

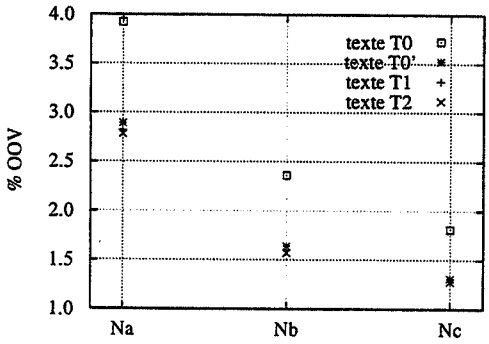


Figure 1: Taux de MHV sur le corpus de développement $\mathcal{T}$ pour différentes normalisations N_a, N_b, N_c et différents textes d'apprentissage T_0, T_0', T_1, T_2 avec un vocabulaire de 65k mots.

[2]Ce texte constitue la ressource textuelle de base fournie à tous les partenaires de l'ARC B1

Table 2: Nombre de 2-grammes et de 3-grammes dans le ML, et perplexité pour différents textes d'apprentissage *LeM*, *LeM + MD*, *LeM + MD + AFP*.

	LeM	*LeM + MD*	*LeM+ MD + AFP*
nb mots	185 M	191 M	255 M
nb bg	11.9 M	12.1 M	13.5 M
nb tg	13.9 M	14.3 M	18.1 M
ppx.	137.7	137.3	135.2

Sélection des mots de vocabulaire Une méthode classique de sélection d'un vocabulaire de taille N, consiste à prendre les N mots les plus fréquents d'un texte représentatif de la tâche de reconnaissance. La sélection du texte le plus représentatif conduit alors à une liste de mots ayant une meilleure couverture, comme l'illustre la figure 1 : les résultats montrent que la couverture obtenue à partir de textes de taille différentes 40M (T_0'), 105M (T_2) et 185M (T_1) est comparable, mais que l'utilisation de textes plus anciens (T_0) conduit à une dégradation nette de la couverture lexicale. Ce résultat conduit à la constatation que la date du texte est plus importante que la taille du texte. Seuls des gains mineurs peuvent encore être obtenus si l'on choisit un texte récent de grande taille (T_2) à un texte plus grand encore (T_1), mais contenant d'autres textes plus anciens.

3.2. *Modèles de langage*

La grande variété lexicale du français implique non seulement des problèmes de couverture lexicale, mais également une estimation plus difficile des modèles de langage. L'utilisation de textes de très grande taille est alors primordiale. Par rapport au système évalué dans SQALE où le texte d'apprentissage avait une taille de 40M de mots, nous avons accru significativement la quantité de texte; nous avons utilisé dans nos expériences 255 M de mots :
LeM: 185 M de mots du journal *Le Monde* années 87-96,
MD: 6 M de mots du journal *Le Monde Diplomatique*, années 89-96,
AFP: 64 M de mots de l'*Agence France Presse*, années 94-96.
Dans la table 2, nous avons indiqué la taille des modèles 3-gramme en fonction des textes utilisés. Tous les modèles de langage présentés ont été obtenus avec une technique dite de repli (*"Back-off"*) de Good-Turing. Dans ces modèles, les 3-grammes singletons ont été exclus.

Afin de limiter plus encore l'effet de cette mauvaise estimation, nous avons interpolé le modèle 3-gramme de mots avec un modèle 2-gramme de classes [Jar96], qui, par une plus grande généralisation, gère mieux ce phénomène. La formule d'interpolation utilisée est la suivante :

$$P_{int}(w) = \lambda P_{tg}(w) + (1-\lambda) P_{bc}(w)$$

où $P_{int}(w), P_{tg}(w)$ et $P_{bc}(w)$ sont respectivement la probabilité interpolée de la séquence de mots, la probabilité obtenue par le modèle 3-gramme de mots et la probabilité obtenue par le modèle 2-gramme de classes. Une valeur optimale de $\lambda = 0.9$ a été obtenu expérimentalement. Cette interpolation conduit à une légère baisse de la perplexité (135 à 131) mais également à une baisse relative du taux d'erreur de 1,5%, en réduisant les fautes d'accord à court terme.

4. RÉSULTATS EXPÉRIMENTAUX

Les résultats présentés ici sont ceux obtenus par le système évalué dans le cadre de l'AUPELF.

4.1. *Le système* AUPELF

Le système de reconnaissance est décrit dans les détails dans [Add97b]. Nous en donnons ici uniquement les grandes lignes.

modèles acoustiques 39 paramètres cepstraux (y compris les dérivées du 1er et du 2ème ordre) constituent les paramètres acoustiques. Ils ont été obtenus à partir du spectre Mel estimé sur une bande de fréquence de 8 kHz. Chaque modèle acoustique est un CDHMM à 3 états, représentant un phone en contexte. Des modèles séparés pour les hommes et les femmes ont été construits. Les modèles ont été estimés sur un ensemble de 66 500 phrases prononcées par 120 locuteurs (BREF[Lam91]).

modèle de langage Nous utilisons des modèles 2 et 3-gramme utilisant une liste de 65k mots, estimés sur les 255 M de mots décrits précédemment, en excluant les 3-grammes singletons et en lissant à l'aide d'une technique de repli de Good-Turing.

lexique Les lexiques d'apprentissage et de reconnaissance ont été développés au LIMSI à partir d'une phonétisation automatique [Bou96]. Chaque entrée lexicale a été transcrite phonétiquement, à l'aide d'un ensemble de 34 phones (incluant le silence). Un graphe de prononciation est associé à chaque mot, afin de prendre en compte les variantes de prononciation.

décodeur Le décodage s'effectue en 3 passes. Lors de la première passe, nous générons un graphe de mots en utilisant un ML 2-gramme (2,2 M de 2-grammes) et 3000 modèles de triphones dépendants de la position dans le mot, et 8000 états partagés. La seconde passe filtre le graphe à l'aide d'un ML 3-gramme (14 M de 3-grammes, 22 M de 2-grammes), et des modèles de triphones indépendants de la position dans le mot (9000 états partagés distribués sur 5000 modèles). La troisième passe est précédée par une adaptation non supervisée, fondée sur la méthode MLLR [Leg95], en utilisant les hypothèses produites lors de la deuxième passe. Le ML interpolé évoqué précédemment n'est utilisé que dans cette passe.

résultats AUPELF Les résultats obtenus sur les corpus $\mathcal{T}$ de développement et de test, à l'aide du système 65k, sont représentés dans la table 3. Le corpus de développement était rendu plus difficile par la présence d'un locuteur dont la langue maternelle n'était pas le français. Le système du LIMSI a obtenu le meilleur résultat officiel dans cette catégorie, avec un taux d'erreur de 11,2%.

Table 3: Résultats officiels obtenus sur les corpus $\mathcal{T}$, en utilisant un lexique 65k et un ML interpolé.

test	*dev.* $\mathcal{T}$	*eval.* $\mathcal{T}$
65k	12,7	11,2

4.2. *Résultats 20k et 65k*

Une comparaison des résultats obtenus à l'aide des systèmes 20k et 65k est résumée dans la table 4.
2 systèmes 20k ont été testés: le $\widehat{20k}$ utilise un ML estimé sur 255 M de mots, alors que le ML du 20k est estimé sur le corpus T_0 (40 M de mots). Le gain observé en augmentant ainsi la taille du texte est faible : en réduction relative du taux d'erreur, le gain est de 9% sur $\mathcal{T}$ et 16% sur $\mathcal{T}'$. Au contraire, on peut observer que le gain obtenu en passant du système 20k au système 65k est important : 40% de réduction relative du taux d'erreur sur $\mathcal{T}$ et $\mathcal{T}'$. Ce gain s'explique par une synergie des gains aux niveaux lexical, syntaxique et acoustique. Il illustre ainsi l'importance que l'on doit apporter à disposer d'une couverture lexicale suffisante, pour espérer améliorer les résultats de reconnaissance par l'amélioration du ML.

Table 4: Résultats de reconnaissance obtenus par les systèmes 20k et 65k sur les corpus $\mathcal{T}$ et $\mathcal{T}'$

$\mathcal{T}$	*%MHV*	*%err.*	$\mathcal{T}'$	*%MHV*	*%err.*
20k	6,38%	23,85%	20k	3,6%	17,3%
$\widehat{20k}$	6,4%	21,8%	$\widehat{20k}$	3,6%	14,6%
65k	1,3%	12,9%	65k	0,5%	8,9%

4.3. *Contrôle du taux MHV*

Nous pouvons mesurer l'impact sur le taux d'erreur du contrôle artificiel du taux de MHV, en nous limitant au système 65k. Nous pouvons observer dans la table 4 une réduction du taux d'erreur de plus de 4 fois la réduction du taux MHV (-4% de taux d'erreur alors que la réduction du taux MHV est de 0,9%) lorsque l'on passe du corpus $\mathcal{T}$ au corpus $\mathcal{T}'$. Si l'on complète cette observation par la mesure de perplexité sur ces 2 corpus (135 sur $\mathcal{T}$ et 106 sur $\mathcal{T}'$, soit une réduction de 20%), nous pouvons en déduire que le contrôle du taux MHV conduit à un filtrage des phrases difficiles. Le corpus $\mathcal{T}'$ est donc non seulement mieux couvert par le vocabulaire 65k mais également mieux modélisé par le ML.

4.4. *Analyse des erreurs*

Nous pouvons attribuer une grande part des erreurs à la faiblesse du modèle de langage. Cette assertion repose sur notre analyse manuelle de la typologie des erreurs observées, mais est également étayée par l'analyse des erreurs en fonction de la fréquence des mots concernés.

L'analyse manuelle des erreurs de reconnaissance nous apprend qu'une grande proportion de celles-ci provient d'erreurs d'accord à court terme en genre, nombre ou personne; environ 40% des erreurs de confusion se sont portées sur les problèmes d'homophonie, erreurs pour lesquelles seul le ML est responsable. 15% des erreurs de substitution sont causés par la présence de noms propres qui sont difficiles à modéliser, à la fois au niveau syntaxique, lexicale et acoustique : les noms propres sont souvent des mots rares, dont la fréquence évolue rapidement dans le temps, et qui (en particulier pour les noms propres étrangers) conduisent à de nombreuses prononciations plus ou moins prévisibles.

Pour mieux comprendre la manière dont le taux d'erreur peut être lié aux capacités du modèle de langage, il nous a semblé intéressant de remplacer le moyennage du taux d'erreur par phrase, par un moyennage par classe de fréquence de mot. A cette fin, le vocabulaire a été partitionné en I classes de fréquence $]K_{i-1}, K_i]$, où les rangs $K_i, i = 1, \ldots, I$ sont distribués logarithmiquement le long de l'axe des rangs de fréquence de mots. Chaque mot w_n du corpus de test, dont la fréquence observée sur le

corpus d'apprentissage conduit au rang de fréquence k_n, est associé avec la région K_i, telle que $k_n \in]K_{i-1}, K_i]$. L'ensemble des mots hors vocabulaires présents dans le corpus de test (0,45%) ont été associés à une classe de rang $> 65k$ (fréquence 0). Le résultat de cette analyse du taux d'erreur par régions de fréquence est visualisé dans la figure 2.

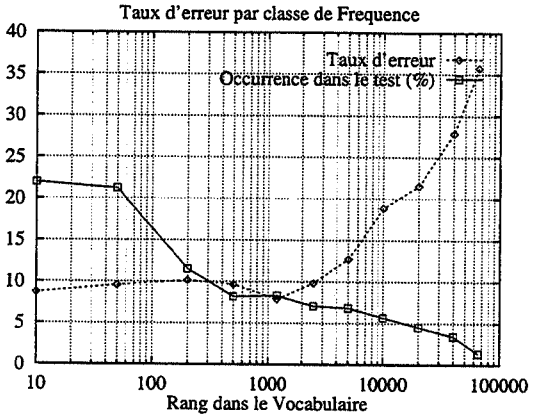


Figure 2: Taux d'erreur de mots et fréquence d'occurrence des mots du test, en fonction des régions de fréquence définies sur le vocabulaire 65k. Chaque point correspond à la limite supérieure de la région. 12 régions ont été définies et distribuées logarithmiquement entre 0 et 65000; la classe de rang $> 65k$, correspondant à 100% d'erreur n'a pas été représentée.

Nous pouvons observer sur la figure 2 que le taux d'erreur pour les rangs $k_n > 5000$ augmente de manière très importante, mais que cela ne concerne que 15% du corpus, c'est-à-dire la partie du texte non couverte par les 5000 mots les plus fréquents (les 7 premières régions). La première région contient les 10 mots les plus fréquents *de, la, l', le, à, et, les, des, d', un*, qui sont des mots très courts, voire monophones homophones, qui sont donc acoustiquement difficiles à discriminer. Les meilleurs résultats sont obtenus pour les mots de la 5^e région (rang entre 500 et 1200) : ces mots sont bien représentés dans le texte d'apprentissage, en général polysyllabiques, donc favorisés à la fois par la syntaxe et l'acoustique.

Seules des erreurs marginales peuvent être imputées à une phonétisation, comme par exemple des liaisons ou des e muets manquants dans le dictionnaire de prononciation. Cependant, on ne peut pas sous-estimer les progrès potentiels que nous pouvons réaliser grâce à une meilleure modélisation acoustique, ne serait-ce que pour expérimenter différentes pondérations entre les modèles acoustiques et de langage dans le décodeur.

5. Conclusion

Nous résumons ici les principales observations que nous avons pu accumuler lors des expériences décrites.
Des gains modestes en couverture lexicale peuvent être acquis en optimisant la normalisation et la sélection des textes. Pour cette dernière la date du texte semble être plus importante que sa taille. L'acroissement de la taille du texte d'apprentissage de 40 à 255 M de mots a permis une réduction significative de la perplexité et du taux d'erreur. L'interpolation du modèle 3-gramme de mots avec un 2-gramme de classes a conduit à un faible gain. L'utilisation d'un vocabulaire de 65k en lieu et place d'un vocabulaire 20k permet une réduction relative du taux d'erreur de 40%, ce qui peut être expliqué par la synergie d'améliorations simultanées aux niveaux lexical, syntaxique et marginalement acoustique. Nous avons montré expérimentalement que le contrôle artificiel du taux de MHV conduit à une contrainte sur la perplexité, ces 2 contrôles induisant une forte amélioration des résultats de reconnaissance. La plupart des erreurs de reconnaissance sont dues à des homophones, parmi lesquels réside une grande proportion de variantes en genre et en nombre. Enfin, nous avons mis en évidence que le taux d'erreur sur les mots les moins fréquents augmentait de manière très importante.
Une conclusion générique à l'ensemble de ces observations peut être que l'amélioration des techniques de modélisation linguistique est la voie de nous nous devons d'explorer dans les années à venir, en reconnaisance de la parole pour la langue française.

6. Remerciements

Une partie de ce travail a été effectuée dans le cadre des Actions de Recherche Concertée "Linguistique, Informatique et Corpus Oraux", financés par l'AUPELF-UREF.

Bibliographie

[Add97a] Adda G. et al., *Text Normalization and Speech Recognition in French* EuroSpeech'97, Rhodos, septembre 1997.

[Add97b] Adda G. et al., *Le système de dictée du LIMSI pour l'évaluation* AUPELF'97 1ères JST FRANCIL, Avignon, avril 1997.

[Bou96] Boula de Mareüil P. *Pour une approche par règles en transcription graphème-phonème* Séminaire GDR-PRC CHM Lexique et Communication Parlée, Toulouse'96 France.

[Dol97] Dolmazon J.M. et al., *ARC B1 - Organisation de la 1e campagne* AUPELF *pour l'évaluation des systèmes de dictée vocale* 1ères JST FRANCIL, Avignon, avril 1997.

[Gau94] Gauvain J.-L. et al., *Speaker-independent continuous speech dictation* Speech Communication **15**, pp. 21-37, septembre 1994.

[Jar96] M. Jardino M. *Multilingual stochastic n-gram class language models* IEEE ICASSP-96, Atlanta, 1996.

[Lam95] Lamel L. et al., *Issues in Large Vocabulary, Multilingual Speech Recognition* Eurospeech'95, Madrid, septembre 1995.

[Lam91] Lamel L. et al., *BREF, a Large Vocabulary Spoken Corpus for French* EuroSpeech'91, Genoa, septembre 1991.

[Leg95] C.J. Legetter C.J., Woodland P.C. *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models* Computer Speech & Language, **9**, pp. 171-185, 1995.

[You97] Young S.J. et al. *Multilingual large vocabulary speech recognition: the European* SQALE *project* Computer Speech & Language, **11**(1), pp. 73-89, janvier 1997.

Etude de la recombinaison de plusieurs classifieurs appliquée à deux tâches de reconnaissance de la parole

CERISARA Christophe, AFIFY Mohammed, HATON Jean-Paul

LORIA
BP 239, 54506 VANDOEUVRE-les-NANCY Cedex
FRANCE
email: {cerisara, afify, jph}@loria.fr

ABSTRACT

This article deals with two different recombination models for multi-band speech processing. The first one is a linear model: the scores returned by each classifier are weighted and summed to compute the final score for each class. The weights are learned with a Minimum Classification Error algorithm. The second one is a MLP trained with classical back-propagation algorithm. We have tested these two models in a phonetic recognition task and in a language identification task. Results of these experiments show an enhancement of the classical models for these two tasks as well as the possibility to extend this principle to several speech related tasks.

1. INTRODUCTION

Dernièrement, plusieurs travaux de recherche utilisant une approche multi-bandes pour des tâches de reconnaissance phonétique de la parole [Bou97] et d'identification du locuteur [Bes98] ont été menés. Le travail présenté dans cet article étudie tout particulièrement la phase de recombinaison dans de tels systèmes, car il nous a semblé que cette phase était particulièrement critique pour ceux-ci. Nous avons voulu, de plus, tester le modèle pour deux tâches de reconnaissance de la parole, la reconnaissance phonétique et l'identification du langage, afin de comparer le gain apporté par le modèle multi-bandes à ces différentes tâches.

Le modèle multi-bandes s'appuie sur des recherches psycho-acoustiques menées par Fletcher et reconsidérées dans [All94]. Ce modèle consiste à filtrer le signal acoustique en plusieurs « bandes » fréquentielles, à traiter chacune d'elle par un système de reconnaissance indépendant, puis à recombiner les différents résultats au moyen d'un classifieur. C'est ce module de recombinaison que nous allons étudier dans cet article.

2. DÉFINITION DES MODULES DE RECOMBINAISON

2.1 Recombinaison linéaire

La recombinaison de plusieurs classifieurs peut être réalisée linéairement. Nous partons de l'hypothèse que les classifieurs retournent, pour chaque élément du corpus d'apprentissage ou de test, un « score » associé à chaque classe. Les classes considérées peuvent par exemple correspondre aux phonèmes du français et le score à la probabilité qu'un segment acoustique représente un tel phonème. Ainsi, une recombinaison linéaire consiste tout simplement à choisir le score minimum ou maximum parmi les « nouveaux » scores suivants :

$$sc(i) = \sum_{bd} \alpha_{bd,i} sc(i,bd)$$

où *sc(i)* représente le score final de la classe *i*, *sc(i,bd)* représente le score retourné par le classifieur *bd* pour la classe *i*, et $\alpha_{bd,i}$ représente un coefficient à estimer.

Nous avons utilisé deux catégories principales de recombinaison linéaire:

Le modèle linéaire simple Dans ce modèle, les coefficients $\alpha_{bd,i}$ sont choisis de manière empirique. Nous avons testé par exemple le cas où tous les coefficients sont égaux, ce qui revient à calculer la moyenne des scores retournés par tous les classifieurs.

Le modèle MCE Nous appelons modèle MCE un modèle linéaire dont les coefficients sont entraînés grâce à l'algorithme d'apprentissage dit de Minimisation de l'Erreur de Classification (*Minimum Classification Error*).

L'avantage principal de la recombinaison linéaire, mis à part sa simplicité, est d'être extrêmement rapide, et d'un coût négligeable face à celui des classifieurs.

2.2 Algorithme MCE

L'erreur de classification n'est pas directement minimisable par les algorithmes classiques de rétro-propagation du gradient, car cette erreur n'est pas dérivable. L'algorithme MCE construit une fonction de coût dérivable approchant l'erreur de classification. La fonction de coût que nous avons choisie est la fonction sigmoïde suivante:

$$l_{bon}(x) = \frac{1}{1 + e^{-\lambda d_{bon}(x)}}$$

où $l_{bon}(x)$ est la fonction de coût associée à l'exemple x qui appartient à la classe *bon*. $d_{bon}(x)$ est une fonction qui « mesure » l'erreur de classification de x par rapport à toutes les classes possibles. Intuitivement, si nous utilisons le minimum de la distance de x par rapport aux centroïdes de toutes les classes pour classer x, alors plus x sera éloigné de son « bon » centroïde et plus il sera proche des « mauvais » centroïdes, plus $d_{bon}(x)$ sera grand.

Classiquement, la fonction utilisée est la suivante:

$$d_{bon}(x) = -g_{bon}(x) + \left[\frac{1}{M-1}\sum_{j,j\neq bon} g_j(x)^{\eta}\right]^{1/\eta}$$

où j parcours l'ensemble des classes et $g_j(x)$ représente le score de l'association de x à la classe j. Il peut s'agir par exemple de l'opposé de la distance de x au centroïde de la classe j. Lorsque η tends vers l'infini, le terme entre crochets tends vers $g_k(x)$ où k est la meilleure classe différente de la bonne classe de x.

Puisque $l_{bon}(x)$ est dérivable, nous pouvons appliquer une méthode classique de rétro-propagation du gradient de l'erreur dans le but de minimiser $l_{bon}(x)$. En pratique, nous modifions après chaque exemple du corpus d'apprentissage, les coefficients $\alpha_{bd,i}$ de la manière suivante:

$$\alpha_{bd,i}(t+1) = \alpha_{bd,i}(t) - \varepsilon\frac{\partial l_{bon}(x)}{\partial \alpha_{bd,i}}$$

Cet algorithme est détaillé dans [Jua97].

2.3 Recombinaison non linéaire

La deuxième méthode que nous avons utilisée pour recombiner les classifieurs est une méthode non linéaire basée sur un réseau de neurones. Nous avons choisi un perceptron multi-couches (PMC) bien adapté à l'apprentissage d'une fonction de décision.

3. DÉFINITION DES TÂCHES DE RECONNAISSANCE DE LA PAROLE

3.1 Principe de base

Nous allons appliquer le même principe, à savoir recombiner les décisions de plusieurs classifieurs à l'aide des modèles présentés ci-dessus, aux deux tâches présentées dans les paragraphes suivants. Les classifieurs utilisés dépendent de chaque tâche, mais ils retournent tous un « score » par classe pour chaque élément du corpus d'apprentissage ou de test.

Le signal de parole est tout d'abord filtré fréquentiellement en quatre bandes, puis, indépendamment pour chaque bande, codé sous la forme de MFCC –Mel Feature Cepstral Coefficients–. Chaque classifieur est ensuite appliqué indépendamment sur chaque bande, et leurs résultats sont alors recombinés par les modèles présentés ci-dessus.

Les domaines fréquentiels des bandes correspondent aux intervalles [0,900 Hz], [800,1600 Hz], [1500,2500 Hz] et [2300, 4000 Hz].

3.2 Reconnaissance phonétique de la parole

La première tâche consiste à reconnaître de la parole en mode *phonèmes isolés*, c'est-à-dire que la segmentation phonétique d'une phrase est donnée au reconnaisseur, qui doit alors décider à quels phonèmes correspondent les segments temporels. Les corpus d'apprentissage et de test –français– sont respectivement, le corpus BREF-80 [Gau91] et le corpus de développement d'Aupelf-Uref [Foh97]. 36 phonèmes de la langue française sont utilisés, et les vecteurs acoustiques sont constitués de 12 coefficients MFCC + Δ + ΔΔ.

Les classifieurs utilisés sont des HMMs du second ordre [Mar97]. Le score qu'ils retournent et qui est passé au module de recombinaison est la log-probabilité que la suite de vecteurs acoustiques appliquée en entrée ait été générée par le modèle du phonème considéré.

3.3 Identification du langage

Cette tâche a pour but d'identifier la langue dans laquelle est parlé un segment donné de parole. Nous avons utilisé le corpus de parole OGI-ML [Mut92] composé d'enregistrements téléphoniques de parole spontanée pour onze langues différentes. Ce corpus étant particulièrement bruité, nous avons utilisé un module d'élimination des segments de bruit, car ceux-ci ne permettent pas de discrimination entre les langues.

Ce module de discrimination bruit/parole utilise deux classes modélisées par des HMMs, l'une pour la parole et l'autre pour le bruit. Ces modèles ont été entraînés sur une petite partie du corpus d'apprentissage dont un étiquetage en classes phonétiques a été réalisé manuellement. Ils sont ensuite utilisés pour segmenter tout le corpus d'apprentissage et de test en segments de parole et de bruit, et seuls les segments de parole sont passés au classifieur principal.

Ce classifieur modélise chaque langue par un simple dictionnaire (*codebook*) obtenu par quantification vectorielle et constitué de 8, 16, 32 ou 64 centroïdes par langue. L'apprentissage de ces modèles est réalisé par l'algorithme LBG.

Cette modélisation d'une langue est très rudimentaire, et de nombreux systèmes plus perfectionnés existent [Rey94][Zis96]. Toutefois, le but que nous nous étions fixé étant de valider l'utilisation du modèle multi-bandes pour une tâche d'identification du langage, nous avons délibérément choisi un modèle qui soit

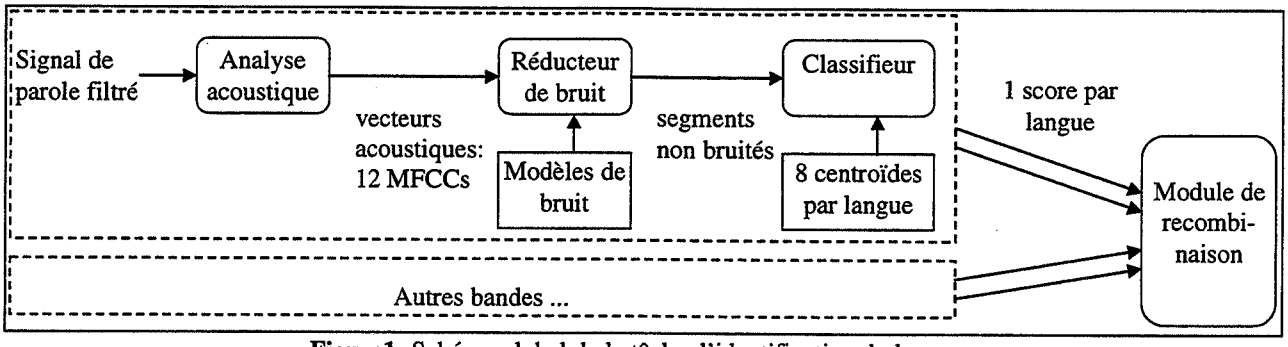


Figure1: Schéma global de la tâche d'identification du langage

facilement manipulable de par sa simplicité, et nous envisageons maintenant d'étendre ce principe à des systèmes plus complexes.

Les classifieurs prennent la décision finale de la langue parlée après avoir traité une phrase de 10 secondes. Ils ont alors calculé un score pour chaque langue, score qui correspond à la somme des distances entre chaque vecteur acoustique et le centroïde le plus proche appartenant au modèle de la langue considérée. Ce sont ces scores qui sont passés au module de recombinaison.

Nous avons testé le système sur deux langues, l'anglais et le japonais, le corpus d'apprentissage total étant composé de 939 phrases de parole spontanée, et le corpus de test total de 252 phrases. Des tests plus complets sont en cours de réalisation et seront disponibles prochainement.

Les vecteurs acoustiques sont composés simplement de 12 coefficients MFCCs. Des expériences plus récentes montrent que l'utilisation conjuguée des coefficients et des Δ-coefficients améliore les résultats, mais nous nous limiterons ici à des vecteurs composés de seulement 12 coefficients. Le système complet est représenté figure 1.

4. RESULTATS DES EXPERIENCES

Dans le cadre des expériences suivantes, le module de recombinaison doit traiter le résultat de cinq classifieurs : les quatre bandes précédemment définies, plus le même classifieur appliqué à l'ensemble du spectre. Intuitivement, nous essayons d'améliorer les performances du classifieur « de base » en lui ajoutant des informations supplémentaires.

4.1 Reconnaissance phonétique

De nombreuses études [Tib97] ont montré que l'un des principal intérêts de l'approche multi-bandes est sa résistance au bruit. Le corpus Bref-80 a été enregistré dans un milieu non bruité, et les résultats qui suivent correspondent à des tests en parole propre. D'autres tests que nous avons effectués en parole bruitée montrent une nette amélioration des résultats. Toutefois, nous nous limiterons dans le cadre de cet article à une étude de la reconnaissance phonétique en milieu non bruité.

Dans la suite, le modèle que nous appellerons *référence* correspond au modèle de base –ici, un HMM2– appliqué à l'ensemble des fréquences utiles. Les taux d'erreurs phonétiques sont donnés dans la Table 1.

Table 1: Tests de reconnaissance phonétique.

	Taux d'erreur
Système de référence	21.1 % (+/- 0.4 %)
Recomb. linéaire avec MCE	20.5 % (+/- 0.4 %)
Recomb. avec MLP	17.5 % (+/- 0.4 %)

Nous pouvons observer une amélioration des résultats du système de référence, surtout grâce à l'utilisation du réseau PMC. Il est également intéressant de noter l'importance de la non-linéarité dans cette tâche de classification.

4.2 Identification du langage

Pour cette tâche, le système de référence est le système composé d'un dictionnaire de N centroïdes par langue, appliqué à l'ensemble des fréquences utiles. Chaque test consiste à identifier la langue parlée parmi les deux suivantes: l'anglais et le japonais. Les résultats de ce système et des classifieurs partiels considérés un par un, sont présentés Table 2.

L'intervalle de confiance de ces résultats est seulement de +/- 4.4 %. C'est peu, mais il est très difficile de faire mieux sur seulement deux langues, car la taille du corpus de test disponible dans OGI n'est pas suffisante.

Table 2: Taux d'erreur pour les classifieurs de base en fonction du nombre de centroïdes par langue.

Classifieurs	8 mix.	16 mix.	32 mix.	64 mix.
Réf.	29.4 %	27.0 %	26.6 %	23.4 %
Bd. 1	27.8 %	33.3 %	29.8 %	32.9 %
Bd. 2	40.4 %	33.7 %	36.1 %	33.7 %
Bd. 3	38.1 %	39.7 %	34.1 %	34.5 %
Bd. 4	39.3 %	34.9 %	36.5 %	35.3 %

Nous avons testé deux types de recombinaison linéaire: La première, la plus simple, consiste à choisir préalablement les valeurs des coefficients. Nous avons testé les cas suivants:

1- Tous les coefficients sont égaux; ceci revient à calculer la moyenne sur l'ensemble des classifieurs. Cette recombinaison sera appelée *moyenne*.

2- Afin de privilégier la réponse du classifieur de

référence, plus riche en information que celle des autres classifieurs, nous avons mis les coefficients des classifieurs partiels à 0.1 et les coefficients du classifieur de référence à 0.6. Cette recombinaison sera appelée *lin2*.

Les résultats des tests pour ce système sont donnés Table 3, pour 8, 16, 32 et 64 centroides par langue.

Table 3: Taux d'erreur pour les classifieurs linéaires simples en fonction du nombre de centroïdes.

Classifieurs	8 mix.	16 mix.	32 mix.	64 mix.
Réf.	29.4 %	27.0 %	26.6 %	23.4 %
moyenne	29.4 %	27.3 %	26.2 %	25.4 %
lin2	31.3 %	26.9 %	25.8 %	23.4 %

Globalement, aucune amélioration n'est apportée par les modèles linéaires simples. Nous avons donc réalisé un apprentissage sur les poids par l'algorithme MCE et par l'intermédiaire du PMC, ce qui a permis d'améliorer les résultats, présentés Table 4.

Table 4: Taux d'erreur pour les classifieurs avec apprentissage des poids (8 centroïdes).

Classifieurs (8 centroïdes)	Taux d'erreur
Système de référence	29.4 %
MCE	21.0 %
PMC	25.4 %

Seul le cas à 8 centroïdes est présenté, car le nombre de coefficients à estimer est directement proportionnel au nombre de centroïdes, et il est difficile d'utiliser les algorithmes complexes comme le MCE ou le PMC lorsque le nombre de paramètres est trop grand. Dans le cas du multi-bandes, la meilleure solution consisterait à réduire le nombre de coefficients MFCC dans chaque bande, tout en augmentant le nombre de centroïdes par bande.

Le module MCE a été entraîné sur un dixième de la base de test et testé sur les neuf autres dixièmes. En ce qui concerne le PMC, les tests ont été réalisés par la technique de « validation croisée », qui consiste à découper la base de test en dix parties, à entraîner le PMC sur neuf dixièmes, puis à effectuer le test sur un dixième. Le score final est la moyenne des scores sur chaque dixième de la base de test.

Conclusions et Perspectives

L'utilisation d'un module de recombinaison, dans le cadre de l'application du multi-bandes à des tâches de traitement automatique de la parole améliore sensiblement les résultats, que ce soit pour la reconnaissance phonétique ou l'identification d'une langue. Toutefois, il est difficile de décider quel type de recombinaison permet d'atteindre les meilleurs résultats dans tous les cas. En effet, si, en théorie, le PMC doit toujours fournir les meilleurs résultats, ce n'est pas systématique en pratique. Ainsi, dans la tâche d'identification du langage, la difficulté du corpus rends l'apprentissage du PMC très délicat, et le minimum global de l'erreur n'est pratiquement jamais atteint. Dans ce cas, le modèle d'apprentissage linéaire réalise de meilleures performances.

Néanmoins, cette étude nous a permis de montrer qu'il était possible, même en milieu non bruité, d'améliorer les résultats des modèles classiques en introduisant une approche multi-bandes aux tâches de parole. De plus, elle nous a également permis de considérer sous un nouveau jour l'importance du module de recombinaison pour de telles tâches ainsi que le problème lié à la dépendance des résultats au corpus.

Bibliographie

[All94] Allen J. B. (1994), "How Do Humans Process and Recognize Speech ?", IEEE Trans. on Speech and Audio Processing, Vol. 2, pp. 567-576.

[Bes98] Besacier L., Bonastre J.-F., Fredouille C. (1998), « Architecture en sous-bandes pour la reconnaissance automatique du locuteur en milieu bruité », RFIA'98, pp. 453-459.

[Bou97] Bourlard H., Dupont S. (1997), « Subband-based speech recognition », ICASSP'97, pp.1251-1254.

[Foh97] Fohr D., Haton J.-P., Mari J.-F., Smaïli K., Zitouni I. (1997), « MAUD: Un prototypt de machine à dicter vocale », 1ères JST 1997 FRANCIL de l'AUPELF-UREF, pp. 25-30.

[Gau91] Gauvain J.-L., Lamel L.-F., Eskénazi M. (1991), « BREF, a large vocabulary spoken corpus for French », EUROSPEECH'91, pp. 505-508.

[Jua97] Juang B.-H., Chou W., Lee C.-H. (1997) "Minimum Classification Error Rate Methods for Speech Recognition ", IEEE Trans. on Speech and Audio Processing, Vol. 5, N° 3, pp. 257-265.

[Mar97] Mari J.-F., Haton J.-P., Kriouile A. (1997), "Automatic word recognition based on second-order hidden Markov models", IEEE Trans. on Speech and Audio Processing., Vol. 5, pp. 22-25.

[Mut92] Muthusamy Y.K., Cole R. A. (1992), « Automatic segmentation and identification of ten languages using telephone speech », Proceedings ICSLP'92, pp. 1007-1010.

[Rey94] Reyes A. A., Seino T., Nakagawa S. (1994), « Three Language Identification Methods based on HMMs », ICSLP'94, pp. 1895-1898

[Tib97] Tibrewala S., Hermansky H. (1997), « Sub-band based recognition of noisy speech », ICASSP'97, pp. 1255-1258.

[Zis96] Zissman M. A. (1996), "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE Trans. on Speech and Audio Processing, Vol. 4, pp. 31-44.

Reconnaissance multi-bandes de la parole bruitée par couplage entre niveaux primitif et d'identification

Hervé Glotin[†‡], Emmanuel Tessier[†], Hervé Bourlard[‡], Frédéric Berthommier[†]

[†] ICP 46 avenue Félix Viallet, 38000 Grenoble, France
[‡] IDIAP Simplon 4, Case Postale 592, 1920 Martigny, Suisse
e-mail: (glotin, tessier, bertho)@icp.inpg.fr, (glotin, bourlard)@idiap.ch

ABSTRACT

We propose to interface primitive and speech recognition stages. The goal is to improve identification of speech corrupted by a high level band-limited noise. We use primitive information linked to the pitch, which primarily allows us to detect the noisy subband. A multi-stream speech recognition system is applied to the selected subbands. We have tested different protocols to select subbands where the speech signal is dominant. One is using a confidence level calculated AFTER entry to a recognition stage. This is based on the Shannon entropy. We extend the use of entropy estimation to detect the corrupted subband from primitive information, BEFORE recognition. This is compared with the more classical R_1/R_0 periodicity estimation. Recognition performances, established on NUMBERS93, are better (error rate at 27.5 %) for BEFORE which is a bottom-up interfacing method, when the lower frequency band is noisy (0 dB SNR).

1. INTRODUCTION

Les modèles de reconnaissance de la parole ont à présent des performances excellentes lorsque le signal n'est pas bruité. Leur fiabilité découle d'une représentation appropriée des signaux, ainsi que de leur redondance. Néanmoins, celle-ci n'est pas exploitée correctement car les performances chutent beaucoup lorsque qu'un signal interférant est ajouté. Afin de lever certains types d'interférences, dites focalisées (bruit de bande étroite ou sinusoïde), l'usage des SBSR ("Reconnaissance de la Parole en Sous-Bandes") a été proposé par Bourlard et coll. [B+96]. Leur principe est de construire une représentation (statistique ou MLP) des signaux de parole dans laquelle un niveau d'indépendance entre des zones fréquentielles est acquis au moment de la phase d'apprentissage. Pour la parole, chacune de ces zones spectrales, dites sous-bandes, code un formant en moyenne. Le but est de limiter la représentation des associations entre régions spectrales éloignées (i.e., de rendre indépendantes les analyses de ces régions, en éliminant l'utilisation des termes de covariance), tout en préservant une propriété de reconnaissance partielle d'une sous-bande ou d'un ensemble de sous-bandes.

Un modèle, dit "de reconnaissance partielle", a été proposé par Green et coll. [G+95], dont la propriété essentielle est de reconnaître des régions présélectionnées dans le plan temps-fréquence, ceci à partir de l'apprentissage en pleine bande. Mais ce modèle n'est applicable (voir [C+96]) qu'à condition de disposer d'un tel sélecteur, puis de restituer en temps réel la représentation utilisée pour la reconnaissance (i.e., en pratique, avec un modèle statistique, d'inverser des matrices de variance covariance partielles). Cette seconde contrainte, de nature algorithmique, n'est pas levée actuellement, mais nous souhaitons développer l'usage du principe de sélection, car il apporte une solution simple au problème du couplage entre le niveau dit primitif et l'étape de reconnaissance, ceci en vue d'acquérir une robustesse vis à vis de signaux interférents de nature variée.

Or, nous disposons de modèles inspirés d'études sur la perception humaine, de type CASA ("*Computational Auditory Scene Analysis*"), capables d'extraire des informations à partir du signal en tenant compte de ses propriétés structurelles selon différentes résolutions temporelles. Les indices primitifs sont : l'harmonicité / périodicité, la modulation d'amplitude (AM), les moments de début et de fin, la différence de temps interaurale (ITD) pour la localisation spatiale. L'extraction de ces indices primitifs repose sur l'utilisation du spectrogramme et de représentations intermédiaires en fonction du facteur de résolution temporelle ou spectrale.

Nous montrerons comment ces algorithmes de type CASA peuvent répondre au principe de sélection (étape de sélection dite AVANT) en prenant comme exemple l'attribut de périodicité ("*pitch*"), extrait par auto-corrélation. Avec ces méthodes, nous isolons les zones du plan temps-fréquence dans lesquelles un signal de parole est dominant sur d'autres signaux interférents (i.e., le rapport signal sur bruit est localement élevé dans le plan temps-fréquence), ou bien, inversement, nous assurons l'identification de la bande la plus bruitée. Puis, nous proposons d'associer une méthode de type CASA et un SBSR qui reconnaît partiellement la parole lorsqu'une seule sous-bande parmi 4 est exclue (i.e., 3 sous-bandes sur 4 sont sélectionnées). Nous nous inspirons d'une étude précédente [TB97] pour présenter une comparaison entre plusieurs modes de sélection, à partir d'indices établis AVANT ou APRES l'entrée dans l'étape de reconnaissance. Pour APRES, nous mettons en œuvre un nouvel indice global proposé par Bourlard et coll., fondé sur la mesure d'entropie. Cet indice permet de choisir la meilleure observation partielle réalisée sur le signal, ceci de façon autonome.

2. ENVIRONNEMENT

Nous utilisons l'environnement STRUT intégrant les étapes de traitement du signal, de décodage acoustico-phonétique et de statistique après correction orthographique. Celui-ci permet aussi de construire des SBSR. Le découpage fréquentiel en 4 sous-bandes est : [0, 901] Hz, [797, 1661] Hz, [1493, 2547] Hz et [2298, 4000] Hz. Pour la reconnaissance, les fenêtres d'analyse du signal sont de 25 ms et se recouvrent sur 12.5 ms. Les MLP du système hybride ANN/HMM [MB95] implémenté dans STRUT génèrent les probabilités des états HMMs (au nombre de 58), ceci pour chaque fenêtre, et l'enchaînement est réalisé par le Viterbi. Les MLP sont entraînés à partir de spectres LPC ("*Linear Predictive Coding*") [H+96] calculés sur chaque sous-bande indépendamment.

Nous notons MLP(xyz) les 4 combinaisons, chacune entraînée sur trois bandes x, y, z, avec pour comparaison un MLP pleine bande (entraîné sur le Log-Rasta PLP). L'une des 4 combinaisons sera sélectionnée pour chaque fenêtre de temps. La phase d'apprentissage et les tests sont réalisés sur NUMBERS93, sans bruitage. Celle-ci est constituée de 2167 phrases téléphonées de nombres produits par 1132 locuteurs. Nous avons utilisé 1534 phrases pour l'entraînement et 384 phrases pour le test. Nous ne réalisons pas d'entraînement à partir de signaux bruités. Le bruit additif est synthétisé à partir d'un bruit blanc Gaussien de bande [0, 430] Hz, donc seule la bande 1 est significativement bruitée sur la base de test. Le rapport signal sur bruit est de 0 dB RMS en moyenne phrase par phrase (silences inclus, et en pleine bande).

3. METHODES

3.1 Principe de sélection

Il est essentiellement fondé sur l'utilisation d'un index permettant de différencier le signal et le bruit. Cet index permet de sélectionner les zones du plan temps-fréquence dans lesquelles le rapport signal sur bruit (RSB) est trop défavorable pour conduire à une reconnaissance. Appliqué sur une représentation intermédiaire du signal AVANT reconnaissance, celui-ci est soit une mesure d'entropie, soit un index de périodicité R_1/R_0. Une mesure d'entropie est appliquée sur les distributions des sorties de chaque MLP (sélecteur dit APRES). L'index entropique est testé dans les deux niveaux afin d'homogénéiser les méthodes de sélection AVANT et APRES.

De façon générale, l'entropie d'un système est une mesure quantitative de son degré de désordre. Appliquée sur une représentation du signal, elle est potentiellement capable de nous indiquer l'existence de structures, par opposition au bruit, dont l'entropie est maximale. L'utilisation de l'entropie sur le spectrogramme a été proposée par [A+97]. Appliquée sur les représentations utilisées durant l'étape de reconnaissance, elle signale si l'information d'entrée s'apparie correctement avec l'information mémorisée, qui décrit l'ensemble des structures devant être reconnues. Le bruit, l'absence de structure, ou bien des distorsions importantes seront aussi diagnostiqués à ce niveau grâce à cet index global. Ainsi, nous pourrons optimiser la reconnaissance en choisissant le MLP(xyz), effectuant une reconnaissance partielle avec le RSB le plus favorable.

3.2. Sélecteur AVANT reconnaissance à partir de l'auto-corrélogramme

Identification de la bande bruitée (IBB) : l'auto-corrélation est évaluée dans chaque sous-bande sur des fenêtres glissantes de 50 ms, toutes les 12.5 ms (l'analyse de périodicité d'un signal de parole réclame des fenêtres temporelles à moyen terme), après démodulation (rectification + passe-bande, selon [BL96]). Nous appliquons une mesure d'entropie pour identifier la sous-bande bruitée. En effet l'auto-corrélation d'un bruit est une distribution quasi-uniforme, donc d'entropie élevée, alors que l'auto-corrélation d'un signal périodique ou harmonique sera plus faible. Nous renormalisons l'autocorrélogramme $ac(\tau)$, afin de calculer l'entropie sur une distribution:

$$H(AC) = -\sum_{t}^{T} \left(\frac{e(t)}{\sum_{t}^{T} e(t)} \right) \log_2 \left(\frac{e(t)}{\sum_{t}^{T} e(t)} \right)$$

avec $e(\tau) = ac(\tau) - \min(ac(\tau)) + 1$ et τ l'axe des délais, T étant le nombre d'échantillons de l'auto-corrélation pris en compte. L'indice R_1/R_0 est le rapport entre (1) R_1=amplitude du premier pic de l'autocorrélogramme dans l'intervalle correspondant aux fréquences [90, 250] Hz et (2) R_0=amplitude en τ=0.

L'IBB est effectuée par comparaison entre sous-bandes, en choisissant celle dont l'entropie d'auto-corrélation est la plus forte ou bien celle dont le rapport R_1/R_0 est le plus faible. Le processus de sélection AVANT consiste à choisir le MLP(xyz) qui ne comprend pas cette sous-bande. A priori, ce modèle sera plus satisfaisant dans les zones voisées, où la différence d'entropie périodicité-bruit est plus forte, alors que dans les zones non voisées, la distinction sera plus difficile (par ex. avec les fricatives).

3.3 Sélecteur APRES l'entrée dans l'étape de reconnaissance

Nous appliquons une mesure d'entropie après normalisation de la distribution des sorties i d'amplitude s_i associées à chaque MLP (dont la fonction de transfert est une sigmoïde). Ces sorties sont assimilables à des probabilités [MB95]. L'entropie permet d'associer une confiance à ce MLP en se basant sur la distribution de toutes ses sorties. Plus la distribution des probabilités est uniforme, plus l'entropie sera grande, moins la confiance pour ce MLP sera grande. Or, l'entropie augmente lorsque du bruit est présent dans l'entrée, car les traits phonétiques sont masqués, et de plus, le bruit lui-même correspond à une réponse du réseau dont l'entropie est forte (ici, le

réseau n'est pas adapté au bruit). Pour une frame Fc, le mode de sélection APRES est défini par le choix du MLP dont l'entropie moyenne sur Fc et ses 2 voisines est la plus faible (fenêtre de 50 ms comme AVANT), car les autres combinaisons sont supposées recevoir le bruit dans leurs entrées. L'entropie est évaluée avec :

$$H(MLP) = -\sum_{i=1}^{N} s_i \log_2(s_i)$$

avec N = nombre d'états de sorties (58).

3.4 Performances d'IBB

Nous établissons les performances d'IBB des différents modes de sélection AVANT et APRES, ainsi que leurs recouvrements (Figure 1). Nous remarquons que AVANT et APRES sont complémentaires (les succès de l'un peuvent correspondent aux erreurs de l'autre). A priori, le sélecteur AVANT, en particulier R_1/R_0, fonctionnerait plutôt lorsque le signal est voisé, alors que le sélecteur APRES serait plus performant dans les segments consonantiques. D'autre part l'ensemble des fenêtres correctement détectées par R_1/R_0 le sont aussi par H(AC) à 5 % près. Les distributions des sélections sont indiquées Table 1 où les erreurs se repartissent d'une façon homogène dans les autres sous-bandes. La Figure 2 illustre la répartition temporelle des IBB. On observe que les deux sélecteurs AVANT ont un comportement similaire pour cet exemple.

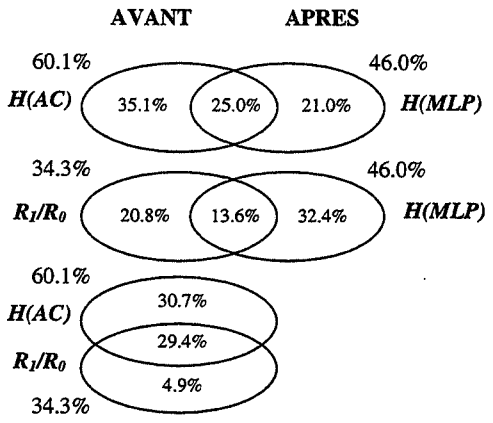


Figure 1 : Taux de succès IBB pour les méthodes AVANT R_1/R_0, H(AC) et APRES H(MLP).

Table 1: Répartition de l'IBB, détaillant le choix de chaque sous-bande sur l'ensemble des fenêtres de temps de la base bruitée (silences compris).

Sous-bande sélectionnée	AVANT *R1/R0*	AVANT *H(AC)*	APRES *H(MLP)*
1	**34.3 %**	**60.1 %**	**46.0 %**
2	20.1 %	15.0 %	15.2 %
3	21.0 %	10.6 %	17.6 %
4	24.6 %	14.4 %	21.2 %

4. Performance de reconnaissance

Le taux d'erreur est la proportion de mots incorrects après décodage acoustico-phonétique et correction orthographique en faisant la somme (délétions + insertions + substitutions). En pleine bande et sans bruit, le taux d'erreur est de 11.3 %. Les résultats de tous les MLP, avec et sans bruit, sont dans la Table 2. Un MLP(xyz) réalise une reconnaissance partielle excluant l'une des sous-bandes parmi 4. Nous vérifions que MLP(234) présente les meilleurs scores, avec 19 % d'erreur, que le signal soit bruité ou non.

Ce taux d'erreur de 19 % est une borne inférieure qui correspond à une identification parfaite de la bande bruitée (toujours la bande 1 dans notre cas). Les écarts observés par rapport à ce score idéal sont imputables aux erreurs d'IBB. Tout d'abord, remarquons qu'une erreur d'IBB se traduit par une double pénalité : la réduction des données et l'introduction de bruit. Dans ce cas, la reconnaissance partielle s'effectue sur 2 bandes seulement (à 0 dB RMS, l'information transmise par la bande 1 est presque complètement masquée) sans que la bande bruitée ne soit exclue. En comparant les Tables 1 et 2, nous pouvons constater que les taux d'erreur obtenus avec les modes AVANT et APRES sont corrélés négativement au taux d'IBB. Néanmoins, les scores obtenus avec R_1/R_0 et H(MLP) sont similaires, ce qui indique une meilleure efficacité de R_1/R_0 peut être liée à une meilleure IBB dans les segments parole+bruit ou bien lorsque les signaux sont voisés.

Table 2: Taux d'erreur de reconnaissance en mots continus suivant le MLP utilisé, sur les phrases test de NUMBERS93. Sélections ad hoc (*), puis sélections AVANT et APRES. Un bruit gaussien de bande [0, 430] Hz à 0 dB est appliqué en sous-bande 1.

MLP sélectionné	Signal **Propre** (*)	Signal **bruité** (*)	Signal **bruité AVANT R_1/R_0**	Signal **bruité AVANT H(AC)**	Signal **bruité APRES H(MLP)**
MLP(123)	12.3 %	50.5 %			
MLP(124)	12.2 %	48.6 %			
MLP(234)	**19.0 %**	**19.2 %**	**37.9 %**	**27.5 %**	**37.6 %**
MLP(134)	14.9 %	55.9 %			
MLP(1234)	**11.3 %**	**55.6 %**			

5. Conclusion et perspectives

Nous exhibons un premier exemple performant de couplage entre un niveau d'analyse primitif et un système de reconnaissance de la parole, testé sur une base de données de référence. Les résultats obtenus avec le sélecteur AVANT sont très prometteurs et ils confirment la possibilité de séparer un signal de parole et une source interférente au cours d'une étape primitive, et selon un mode AVANT ascendant ("bottum-up"). Les autres solutions semblent moins performantes dans le contexte que nous avons choisi mais elles pourraient apporter des solutions lorsque le signal interférant est de nature différente. Pour progresser dans l'élaboration d'un modèle de reconnaissance robuste à des bruits de nature variée, plusieurs méthodes sont envisagées :

- En premier lieu, nous vérifierons ces conclusions en bruitant les bandes 2, 3 et 4. Nos premiers résultats indiquent une différence de comportement des sélecteurs AVANT, avec de meilleures performances de R_1/R_0 par rapport à H(AC), tout en gardant des taux d'IBB satisfaisants.
- Test du même modèle avec des bruits non stationnaires (à bande étroite mais dont la fréquence centrale change). Dans ce but, nous avons limité l'usage d'informations *a priori* sur le bruit (la continuité temporelle n'est pas une contrainte forte, puisque l'IBB est réalisée indépendamment dans chaque fenêtre temporelle).
- Usage d'autres indices primitifs (ITD, AM) de façon à mieux couvrir les zones non voisées, et à améliorer le modèle de sélection AVANT. Remarquons que ces indices sont extraits à partir d'autres représentations intermédiaires exprimées selon des échelles de temps différentes (plus petite pour l'ITD, plus grande pour l'AM).
- Amélioration de l'adaptativité du modèle de reconnaissance de façon à préserver sa capacité de reconnaissance partielle tout en gardant la possibilité de pré-calculer les représentations. Pour cela, nous pourrons tester des combinaisons de sous-bandes plus étroites (mais plus nombreuses) ou qui se superposent fréquentiellement, ainsi que les modèles associatifs à points de recombinaisons syllabiques [MB95].
- Des modes de fusions autre que la sélection / recombinaison de sous-bandes sont envisageables par pondération des probabilités sous-bandes à partir des indices primitifs.

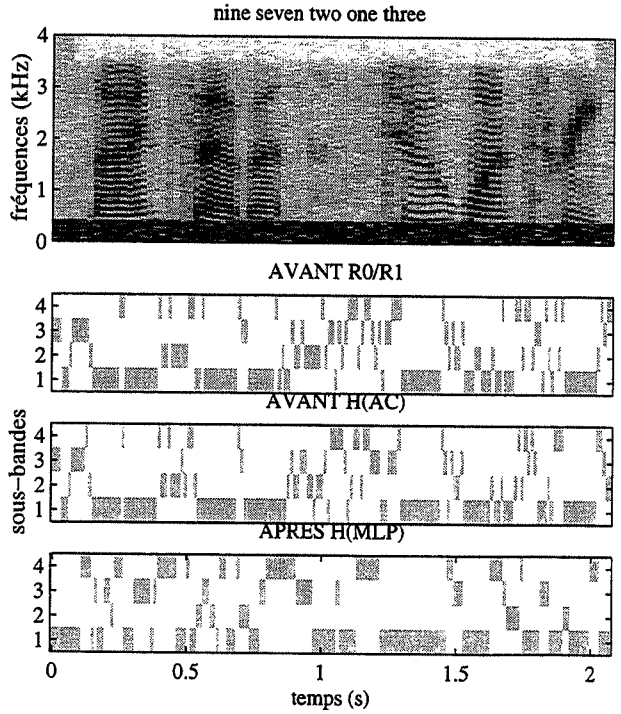


Figure 2 : Analyse d'une séquence bruitée de la base NUMBERS93. De haut en bas : spectrogramme du signal bruité à 0 dB. Dans les 3 figures du bas, pour chaque fenêtre de 25 ms, la bande sélectionnée est indiquée en grisé pour AVANT R_1/R_0, H(AC) et APRES H(MLP).

REMERCIEMENTS

Ce travail a été soutenu par le projet COST249, par Eurodoc, et il entre dans le cadre de SPHEAR, contrat TMR coordonné par Phil Green. Nos remerciements à Stéphane Dupont pour son aide et ses conseils dans l'élaboration du système multi-bandes que nous utilisons pour cet article.

BIBLIOGRAPHIE

[A+97] I. Abdallah, S. Montresor & M. Baudry, "Un algorithme récursif pour la segmentation des signaux de la parole basé sur un critère entropique local", *Actes du 4ième congrès français d'Acoustique,* Marseille, 1997, pp. 85-88

[BL96] F. Berthommier & C. Lorenzi, "Precise and perceptually relevant processing of amplitude-modulation in the auditory system: physiological and functional models", In *Neurobiology*, V, Torre & F. Conti Eds, New-York, Plenum Press, 1996, pp. 139-153

[B+96] H. Bourlard, S. Dupont, H. Hermansky & N. Morgan, "Towards subband-based speech recognition", in *Proc. of European Signal Processing Conference,* Sept. 1996, pp. 1579-1582

[C+96] M. P. Cooke, A. Morris & P. Green , "Recognising occluded speech", in *Proc. of Workshop on the auditory bases of speech processing,* Keele, 1996, pp. 297-300

[G+95] P. D. Green, M. P. Cooke & M. D. Crawford, "Auditory scene analysis and HMM recognition of speech in noise", in *Proc. ICASSP,* 1995, pp. 401-404

[H+96] H. Hermansky, M. Pavel & S. Tibrewala, "Towards ASR using partially corrupted speech", in *Proc. of Intl. Conf. On Spoken Language Processing,* Oct. 1996, pp. 458-461.

[MB95] N. Morgan & H. Bourlard, "Continuous Speech Recognition - An Introduction to the Hybrid HMM/Connectionist Approach", *IEEE Signal Processing Magazine,* Invited Paper, vol 12, n°3, May 1995

[TB97] E. Tessier & F. Berthommier, "A model of the cumulated effect of pitch and interaural delay differences for double vowel segregation", in *Proc. of ICSP'97*, Séoul, 1997, pp. 753-758

Reconnaissance de noms prononcés et épelés au téléphone par modèles de Markov cachés

Frank Neubert, Guillaume Gravier, François Yvon, Gérard Chollet

Ecole Nationale Supérieure des Télécommunications / CNRS-URA 820
46, rue Barrault
75634 Paris Cedex 13, FRANCE

ABSTRACT

A system to retrieve names in a directory from the pronunciation and the spelling is presented, for telephone quality speech in French. The recognizer is based on Hidden Markov modeling of speech units using word models for letters and phone models for name pronunciation. The directory is built automatically from a list of names, each entry consisting in several pronunciation and spelling variants. From the acoustic recognition, the corresponding entry in the directory is found using dynamic alignment of symbol sequences, with insertion, deletion and substitution costs determined from the training data to take into account acoustic confusability. Experimental results show that a mixed recognition strategy combining both spelling and pronunciation improves over the spelling based strategy.

1. INTRODUCTION

La reconnaissance de noms prononcés et épelés représente un enjeu important en reconnaissance automatique de la parole, en raison des applications qui en découlent, par exemple la recherche automatisée d'entrées dans un annuaire. De nombreuses études consacrées à la reconnaissance de l'alphabet ont mis en évidence les difficultés liées à la difficulté de distinguer certaines lettres ("*b*" et "*d*" par exemple..). De plus, lorsque les recherches portent sur la parole transmise au téléphone, ce problème devient bien plus gênant encore. Des expériences ont été faites sur la reconnaissance des alphabets français et anglais, par Cole *et al.* notamment [Col91], [Sch93]. Plus récement, des approches utilisant des modèles de Markov cachés (HMM) des lettres ont été introduites [Jou93], [Jun95]. Ces études, cependant, prennent surtout en compte les connaissances phonétiques au niveau du décodage, sans les réinvestir ensuite dans la recherche lexicale (par alignement dynamique (DTW) des séquences décodées sur les entrées d'un lexique). De plus, le nom lui-même, c'est-à-dire sa prononciation, est rarement employé comme source d'information complémentaire (voir à ce propos [Mey97]).

Cet article présente donc la mise au point d'un système de reconnaissance de noms prononcés et épelés, utilisant des modèles de Markov cachés de phonèmes et de lettres, et une recherche lexicale tenant compte des taux de confusion entre unités phonétiques [Gra97]. Le lexique, généré automatiquement, peut intégrer plusieurs variantes de prononciation et d'épellation pour chaque nom. Après avoir présenté la base de données et le lexique, nous décrirons le décodeur acoustico-phonétique (section 3), puis nous comparerons différentes stratégies en recherche lexicale (section 4). Enfin, après une analyse des causes d'erreurs les plus importantes, des améliorations du système actuel seront envisagées.

2. BASE DE DONNÉES ET LEXIQUE

La base de données Polyphone a été enregistrée grâce à 5 000 appelants, chaque appel incluant la prononciation et l'épellation de trois noms (nom propre, nom commun ou nom de lieu). L'épellation étant libre, le corpus contient, en plus du vocabulaire d'épellation habituel, des comparaisons ("a comme Alain", par exemple), et parfois même des termes aéronautiques ("alpha", "bravo"...). C'est un sous-ensemble contenant 11 920 segments de parole provenant de 3 998 locuteurs que l'on a utilisé pour entraîner et tester le système, après l'avoir divisé en trois parties : le corpus d'apprentissage, contenant 5 390 segments, le corpus de test (6530 segments), et un sous-ensemble de ce dernier, le corpus de test dit "propre", qui contient les 5015 segments dont l'épellation appartient au vocabulaire défini comme standard (excluant les formes d'épellation rares évoquées plus haut); il est employé essentiellement pour évaluer la qualité du décodage.

Le lexique, quant à lui, contient l'ensemble des noms à reconnaître, soit les 8 261 entrées différentes présentées aux appelants. Chaque entrée regroupe une ou plusieurs transcriptions phonétiques du nom prononcé, de même pour l'épellation. Les variantes de prononciation ont été générées par un convertisseur graphèmes-phonèmes développé au cours du projet Onomastica [Fit93], [Yvo96] conçu pour prendre en compte les idiosyncrasies de prononciation des noms propres, ainsi que les particularités d'origine. Quant aux variantes d'épellation, elles sont automatiquement engendrées par un système expert qui envisage pour chaque lettre ou groupe de lettres toutes les épellations possibles (par exemple, la lettre "é" peut être épelée "*e accent aigu*", "*e aigu*" ou "*é*", le groupe "nn" se dit "*deux n*" ou bien "*n n*"...). Par exemple, l'entrée "*tillie*" pourra être prononcée [*t*] [*i*] [*l*] [*i*] ou bien [*t*] [*i*] [*j*] [*i*], et être épelée "*t i deux l i e*", ou "*t i l l i e*". Dans la suite du texte, une entrée e_i du lexique sera représentée par $\{n_{i,j=1,...,N_i}, s_{i,j=1,...,S_i}\}$, où $n_{i,j}$ (resp. $s_{i,j}$) est la j-ème variante de prononciation (resp. d'épellation).

3. DÉCODAGE ACOUSTICO-PHONÉTIQUE

3.1. *Modèles et grammaire*

La première étape de la reconnaissance consiste dans la phase de décodage. Le signal de parole subit d'abord une paramétrisation en 12 coefficients MFCC plus le logarithme de l'énergie, et les dérivées premières et secondes de ces coefficients, soit des vecteurs à 39 composantes, mesurés toutes les 10 ms sur des fenêtres de 25.6 ms. Pour calculer les modèles (phonèmes et lettres) nécessaires à la reconnaissance des noms prononcés et épelés, le corpus d'entraînement subit d'abord une segmentation automatique à partir des transcriptions orthographiques (à l'aide d'un phonétiseur), en effectuant un alignement forcé employant des modèles de phonèmes initiaux (utilisés comme "bootstraps"), après quoi les paramètres des HMM sont réestimés grâce à l'algorithme de Baum-Welch; les 36 modèles de phonèmes ainsi calculés sont des HMM multigaussiennes à trois états. Les modèles des lettres sont ensuite créés en concaténant les modèles de phonèmes correspondant à leur prononciation, puis en effectuant une réestimation de ces modèles en contexte d'épellation. En raison des liaisons existant après certains mots comme "*deux*", certaines lettres ont deux modèles, l'un sans liaison et l'autre avec (par exemple "*a*" et "*za*"). Ainsi, le vocabulaire d'épellation standard, constitué de 41 termes, correspond en tout à 55 modèles.
La grammaire initiale (gram 1), très simple, définit un événement comme une suite de phonèmes avec silence éventuel (correspondant à la partie prononcée), suivie d'une suite de lettres (correspondant à l'épellation), ce que l'on peut schématiser ainsi :
gram1 = $\{phone|sil\}$ $\{lettre\}$. Notons que cela autorise le locuteur à omettre l'une ou l'autre partie (prononciation ou épellation), ce qui est courant dans les situations réelles. Enfin, on a forcé dans la grammaire la reconnaissance d'un silence court après chaque lettre ([Pye94]), ce qui s'est avéré utile pour obtenir une meilleure précision sur les lettres. A partir de la grammaire ainsi définie, le décodeur recherche la séquence de phonèmes et de lettres la plus probable.

3.2. *Résultats*

Dans cette section sont présentés les résultats des performances du décodage pour la reconnaissance des "unités phonétiques" (phonèmes et lettres) ; la table 1 montre ainsi les taux de reconnaissance et l'exactitude obtenus avec la grammaire décrite plus haut (voir 3.1) sur les corpus d'apprentissage et de test propre. La précision correspond au nombre d'unités correctement reconnues diminué du nombre d'insertions, et constitue une mesure de la qualité du décodage bien plus fiable que le taux de reconnaissance total. Comme on le voit, bien que le taux de reconnaissance soit assez bon, il y a encore beaucoup d'insertions.

Afin de réduire le nombre d'insertions, on a introduit une pénalité fixe de transition entre deux modèles (sous la forme du logarithme d'une probabilité) : plus celle-ci est petite, plus le nombre d'insertions diminue, ce qui accroit, simultanément, le nombre de suppressions. La figure 1 montre les taux de reconnaissance et la précision

Table 1: Taux de reconnaissance (en %) et précision pour les phonèmes et les lettres, obtenus avec gram 1 :

	phonèmes		lettres	
	correct.	précis.	correct.	précis.
apprent.	58.71	14.42	83.16	74.56
test propre	57.99	-6.72	81.53	69.72

obtenus sur le corpus d'apprentissage en fonction de cette pénalité de transition (notée P dans la suite), ainsi que la performance globale en recherche lexicale ; on peut remarquer que cette dernière varie peu en fonction de P (voir 4, recherche lexicale) et que le décodage des lettres est plus fiable que celui des phonèmes.

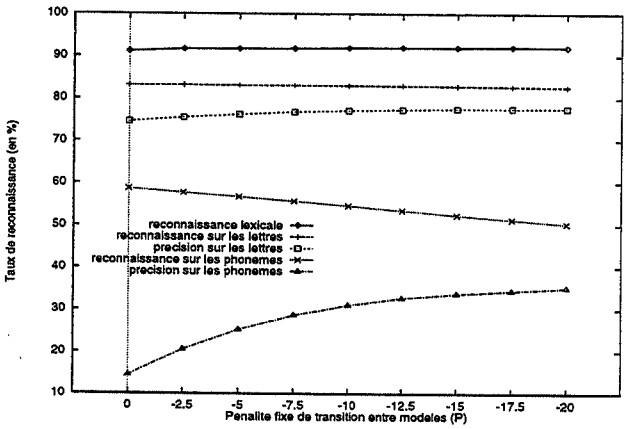


Figure 1: Taux de reconnaissance sur le corpus d'apprentissage en fonction de la pénalité fixe de transition entre modèles.

Le nombre élevé des insertions de phonèmes est également dû à une difficulté de localisation du début de l'épellation par le décodeur, ce qui tend à surévaluer le nombre phonèmes reconnus, et à sous-évaluer le nombre de lettres lors du décodage : ainsi, environ 30% des lettres sont identifiées comme suites de phonèmes (voir 5.).
Par ailleurs, une étude plus détaillée des erreurs faites sur le décodage des unités phonétiques met en évidence les confusions classiques entre lettres acoustiquement proches ("*p*" et "*t*", "*f*" et "*s*", etc..) et entre phonèmes similaires ([*m*] et [*n*] par exemple). Comme mentionné dans de nombreuses études consacrées à la reconnaissance de l'alphabet (voir [Jou93], [Loi96]), les approches par HMM ne sont pas aptes à distinguer entre des unités qui diffèrent seulement par de courts événements transitoires. Par exemple, les lettres "*m*" et "*n*", qui ont en commun la même voyelle initiale, se distinguent seulement par la consonne finale ; d'autre part les phonèmes [*m*] et [*n*], ressemblants, seront souvent confondus, entraînant ainsi une confusion sur les lettres "*m*" et "*n*".

4. RECHERCHE LEXICALE

4.1. *Stratégie de base*

La deuxième étape consiste ensuite à rechercher un nom dans le lexique, à partir du résultat obtenu en sortie du

module de décodage (une séquence de phonèmes et/ou de lettres). Pour ce faire, on calcule une mesure de dissimilarité entre la séquence reconnue r et chaque entrée lexicale e_i :

$$D(r, e_i) = \beta(\min_j d(r_s, s_{i,j})) + (1 - \beta)(\min_j d(r_n, n_{i,j}))$$

où r_s (resp. r_n) est l'épellation (resp. la prononciation) reconnue. L'entrée finalement sélectionnée sera celle qui minimise cette mesure. Le paramètre β permet de pondérer la contribution respective de l'épellation et de la prononciation dans le calcul de D. La mesure de dissimilarité d(,) est calculée par alignement dynamique, en tenant compte de coûts spécifiques pour toutes les substitutions, insertions et suppressions possibles au niveau des unités phonétiques. Ces coûts peuvent être fixés arbitrairement. Cependant, comme certaines paires d'unités sont plus souvent confondues que d'autres, il est naturel de leur assigner un coût plus faible. Ainsi, le coût pondéré pour la substitution de x par y est $-\log(p(y|x))$, où $p(y|x)$ est estimé sur le corpus d'apprentissage par calcul d'une matrice des confusions. La même méthode est employée pour déterminer les coûts d'insertion et de suppression. Cette approche est semblable à la procédure d'alignement à partir des HMM décrite dans [Jou93] : plutôt que d'introduire des connaissances phonétiques au niveau du décodage afin d'améliorer la discrimination entre modèles comme dans [Loi96], cette connaissance dérive des résultats obtenus à l'apprentissage et est réinvestie dans la recherche lexicale et non pour le décodage. Ainsi, avec $\beta = 0.5$ et sans pénalité de transition (P=0), on obtient un taux de reconnaissance lexicale de 52.11% avec des coûts binaires, contre 82.11% avec des coûts pondérés. Pour les exprériences d'accès lexical faites dans la suite, on fixera P égal à -17.5. Les taux de reconnaissance lexicale obtenus avec différentes valeurs de β sur le corpus d'apprentissage sont reportés sur la figure 2, et permettent de déterminer le facteur de pondération optimal : $\beta = 0.575$, ce qui concorde avec la meilleure fiabilité du décodage observée sur l'épellation (voir 3.2).
Avec cette valeur optimale, on obtient un taux d'accès lexical de 79.35% (intervalle de confiance à 95% : [78.36% , 80.33%]) sur le corpus de test complet (contre 21.86% avec $\beta = 0$, et 76.55% lorsque $\beta = 1$). Par ailleurs, si on fixe $\beta = 0.5$, lorsqu'on prend une pénalité de transition nulle (P=0), la valeur obtenue est de 76.66% sur le test, contre 78.94% avec la pénalité optimale (P=-17.5), ce qui montre bien l'utilité de la double optimisation effectuée jusqu'ici (sur la pénalité de transition P, et sur le facteur de pondération β).

4.2. *Présélection*

La recherche lexicale décrite dans la section précédente peut prendre beaucoup de temps, puisqu'il s'agit de calculer les alignements dynamiques de la prononciation et de l'épellation fournies par le décodeur, avec l'ensemble des prononciations et des épellations contenues dans le lexique, comprenant souvent plusieurs variantes pour chaque entrée (voir 2.). Manifestement, cette approche très coûteuse n'est correctement applicable que dans le cas de vocabulaires de moyenne ou de petite taille.
Comme première étape vers une recherche des noms dans de grands lexiques, nous présentons ici une stratégie alternative intéressante. Elle consiste en deux étapes successives : tout d'abord, sur la base de l'épellation seule, un ensemble de N entrées est sélectionné par une recherche lexicale approximative mais très rapide; ensuite, ce sous-ensemble du lexique est exploré entièrement, en appliquant alors la méthode décrite précédemment. La phase de recherche lexicale approximative utilise une variante de l'algorithme proposé dans [Ofl96], où une représentation arborescente du lexique permet de retrouver rapidement les entrées situées à une distance d'édition inférieure à un seuil fixé d d'une entrée donnée. La présélection utilisée ici suit le même principe, à ceci près que l'on recherche à présent un nombre fixe N d'entrées les plus proches. Si cela ralentit un peu la recherche accélérée, le module de présélection reste malgré tout très rapide : le traitement d'une entrée provenant du décodeur est environ 9 fois plus rapide pour une présélection sur les 100 voisins les plus proches, avec seulement une légère diminution du taux de reconnaissance lexicale. Par exemple, pour N=250 voisins présélectionnés sur le corpus de test on obtient un taux de performance de 78.27%, contre 79.09% sans présélection (on a pris P=-7.5 et $\beta = 0.625$).

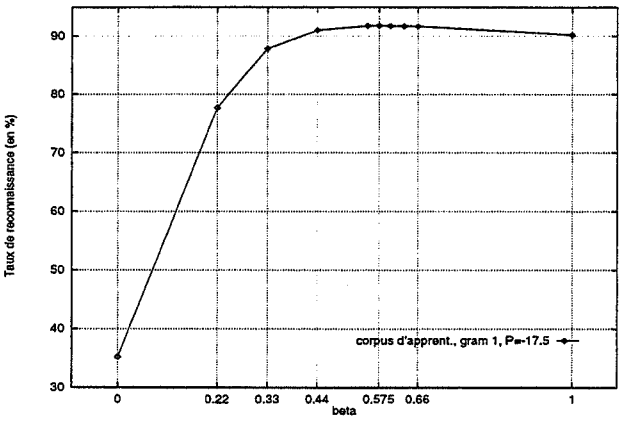


Figure 2: Taux de reconnaissance lexicale en fonction de β (avec P=-17.5 et gram 1).

5. ANALYSE DES ERREURS ET AMÉLIORATIONS

L'analyse des erreurs commises par le système de reconnaissance sur le corpus d'apprentissage (avec les valeurs actuellement optimales de la pénalité de transition entre modèles et de la pondération des distances sur le lexique) a mis en évidence deux points importants :
D'une part, l'existence d'un défaut de localisation du début de l'épellation par le module de DAP (observable sur 60% des erreurs environ), impliquant une surévaluation du nombre de phonèmes réconnus au détriment des lettres. D'autre part, la fréquence de séquences phonétiquement impossibles fournies par le décodeur (par exemple, des suites de plusieurs consonnes consécutives). Pour remédier au premier point, une nouvelle grammaire a été testée (gram 2), dans laquelle on a simplement spécifié la présence d'un modèle de silence entre les phonèmes et les lettres de la séquence reconnue, afin de favoriser la détection du début d'épellation, correspondant alors au silence naturellement marqué par

les locuteurs dans la plupart des cas (si celui-ci est effectivement détecté par le décodeur, ce qui n'est pas toujours le cas, puisque après modification le retard de localisation subsiste encore dans 35% des cas). Cependant, le taux de reconnaissance s'en trouve sensiblement augmenté (voir la figure ci-dessous).

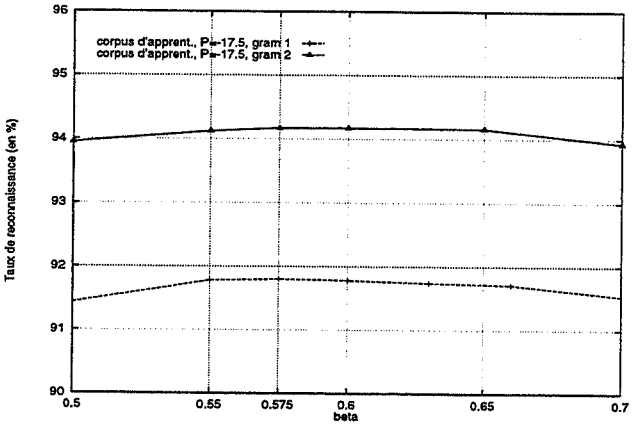


Figure 3: Taux de reconnaissance lexicale en fonction de β, pour gram 1 et gram 2, sur le corpus d'apprentissage.

Grammaires utilisées :
gram1 = $\{phone|sil\}$ $\{lettre\}$
gram2 = $\{phone\}$ sil $\{lettre\}$

Il apparaît ainsi (figure 3) que la valeur optimale pour le facteur de pondération lexicale ne varie pas de manière significative quand on utilise gram 2 au lieu de gram 1 sur le corpus d'apprentissage. Enfin, on notera une amélioration de 3.3 points sur le corpus de test, puisque l'on est passé de 79.35% (avec gram 1 et β optimal), à 82.68% (avec gram 2 et β optimal; l'intervalle de confiance à 95% est alors [81,76% , 83.59%]).
Pour le second point, l'intégration d'un modèle de langage bigrammes est prévue au cours de la phase de décodage, afin de réduire simultanément la probabilité des suites de phonèmes et de lettres impossibles et celle des termes d'épellation hors contexte... sans pour autant imposer une grammaire trop rigide, ce qui s'est avéré inefficace par le passé [Ver97]. Par la suite, d'autres améliorations pourront être envisagées, notamment :
- Reconnaissance forcée avec grammaire incluant les 10 meilleures hypothèses pour détecter la plus probable.
- Détection puis traitements des épellations non conformes.
- Enrichissement des variantes de prononciation du lexique engendrées par le phonétiseur.

6. Conclusion

Les expériences menées sur la reconnaissance des lettres à l'aide de HMM montrent que cette technique ne permet pas d'éviter les confusions classiques entre lettres acoustiquement similaires. Cependant, la connaissance des taux de confusion au niveau du décodage permet d'adapter en conséquence les coûts définis pour l'alignement dynamique.De plus, un algorithme de présélection des entrées lexicales les plus proches permet d'accélérer considérablement la recherche lexicale, avec une très faible diminution des performances.
Par ailleurs, des perfectionnements en cours d'élaboration au niveau du module de décodage (afinage de la détection du début d'épellation, modèles de langage "bigrammes", stratégies "N-Best") devraient permettre d'accroître encore notablement les performances du système.

Remerciements

La société Swisscom finance en partie les travaux décrits dans cet article.

Bibliographie

[Col91] Ronald Cole, Krist Roginski et Mark Fanty. *English alphabet recognition with telephone speech.* EuroSpeech, pp. 479-482, 1991.

[Sch93] P. Schmid *et al.* *Real-time, neural network-based, French alphabet recognition with telephone speech.* EuroSpeech, pp. 1723-1726, 1993.

[Jou93] D. Jouvet *et al.* *Speaker-independent spelling recognition over the telephone.* Int. Conf. on ASSP, pp. 235-238, 1993.

[Jun95] Jean-Claude Junqua *et al.* *An n-best strategy, dynamic grammars and selectively traine neural networks for real-time recognition of continuously spelled names over the telephone.* Int. Conf. on ASSP, pp. 852-855, 1995.

[Mey97] Michael Meyer et Hermann Hild. *Recognition of spoken and spelled proper names.* EuroSpeech, pp.1579-1582, 1997.

[Gra97] G. Gravier *et al.* *Directory name retrieval using HMM modeling and robust lexical access.* Workshop on Automatic Speech Recognition and Understanding, pp. 558-565, 1997.

[Fit93] S. Fitt, M. Schmidt, C. Scott, et M. Jack. *Phonetic transcription standards for European names (ONOMASTICA).* EuroSpeech, Vol. 1, pp. 279-282, 1993.

[Yvo96] François Yvon. *Prononcer par analogie : motivation, formalisation et évaluation.* Ecole Nationale Supérieure des Télécommunications, 1996.

[Pye94] David Pye. *Automatic recognition of continuous spelled Swiss-German letters.* IDIAP, 1994.

[Loi96] Philipos C. Loizou et Andreas S. Spanias. *High-performance alphabet recognition.* IEEE Trans. on Speech and Audio Processing, Vol. 4, pp. 430-445, 1996.

[Ofl96] Kemal Oflazer. *Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction.* Computational Linguistics, Vol. 22, No 1, pp. 73-89, 1996.

[Ver97] Claire Vermonet et Guillaume Etorre. *Reconnaissance de mots épelés indépendamment du locuteur par modèles de Markov cachés.* ENST, 1997.

Indexation vocale à vocabulaire illimité à base de décodage phonétique

S. PEILLON A. FERRIEUX

France Télécom C.N.E.T. - lab. DIH/DIPS
2, av. Pierre Marzin, 22307 LANNION Cedex
FRANCE
e-mail : {stephane.peillon,alexandre.ferrieux}@cnet.francetelecom.fr

ABSTRACT

This paper describes a new speaker and vocabulary independent wordspotting algorithm suited for spoken information retrieval. The retrieval is very fast, and based on a single pronunciation of the word to retrieve. The speech data are first phonetically indexed. Next, two distinct Markov models can perform the wordspotting. One model simply measures an edition distance between phone sequences, whereas the other one is closer to the human way of aligning these sequences. We estimate a 6.1 % error rate for isolated words' discrimination. Discrete distribution training is performed by corrective training and by genetic optimization. Furthermore, we develop an analytical procedure to calculate distribution with a phonetically labelled database.

1. INTRODUCTION

L'arrivée massive dans le grand public de puissances de calcul et de stockage d'informations amène les programmeurs à développer des outils de tri et d'extraction de documents vocaux. Outre le problème persistant de l'interprétation d'une requête en langage naturel, s'ajoute celui de la détection d'une information dans un corpus de parole. Le système présenté ici consiste à indexer une base de documents sonores de façon à pouvoir retrouver un document à partir de son seul contenu et d'une requête vocale.

Un système d'indexation comporte deux parties distinctes : le moteur d'indexation (off-line), et l'outil de recherche (on-line). Si le temps d'exécution du moteur n'est pas fondamentalement critique, il est crucial que la partie on-line soit effectuée bien plus vite que le temps-réel. Le challenge technique consiste donc à transférer le maximum de calculs de la partie on-line vers la partie off-line.

En terme d'applications, on distingue deux approches, selon la liberté d'emploi envisagée : vocabulaire fini, ou vocabulaire illimité. La première permet de profiter des méthodes usuelles de détection de mots clés, mais en hérite aussi certaines limitations comme le manque de souplesse dû à l'apprentissage d'un vocabulaire spécifique. La seconde approche, moins balisée, ouvre la voie à une ergonomie d'un autre ordre de grandeur. Cet article expose cette seconde approche, dans un cadre multi-locuteurs et avec requête vocale.

Dans la littérature, la détection de mots clés commence avec les travaux de [Hig85]. Ces derniers effectuent une DTW des trames acoustiques des mots clés avec celles de la base de données à traiter. Des « références poubelles » concaténées entre elles absorbent la parole de fond. L'inconvénient de la méthode est l'importante masse de calculs à entreprendre.

On trouve par la suite le système de [Ros90] conçu pour reconnaître un grand nombre de mots clés indépendamment du locuteur. Les mots clés sont découpés en unités de taille inférieure au mot dont chacune est modélisée par 3 états d'un HMM. La parole « de fond » est modélisée par des modèles poubelles de mots, de triphones ou de phonèmes. Ce *spotter* à grammaire nulle obtient un FOM[1] d'environ 60 % sur la base de données Stonehenge. Il est en général difficile d'obtenir un bon modèle de rejet dans les HMM-spotter et la fixation d'un seuil de rejet n'est pas aisée. De plus, l'apprentissage des modèles de mots limite fortement la souplesse du *spotter*.

On trouve cependant un article plus proche de notre algorithme chez [Jam94]. Ce dernier propose un système de pré-décodage de la base de données, mais sous la forme d'un treillis de phonèmes (incluant bornes temporelles et vraisemblance). La recherche du mot s'effectue ensuite dans ce treillis à partir d'une transcription phonétique manuelle de la clé de recherche. C'est une approche indépendante du vocabulaire, mais encore un peu coûteuse en calculs. Le FOM est de 67 % sur la base DARPA 'Resource Management'.

Notre document va tout d'abord présenter le principe de notre système avant de développer deux modèles de détection de mots dont l'évaluation est faite dans un premier temps sur une base de mots isolés. On chiffrera les performances en mode *spotting* sur un corpus de phrases au chapitre 5.

2. MÉTHODE UTILISÉE

Le système de détection de mots que nous proposons possède les propriétés suivantes : indépendance au locuteur, vocabulaire illimité, recherche très rapide des informa-

[1] FOM : Figure Of Merit, défini comme étant la moyenne du taux de détection pour un « flux » de fausses alarmes compris entre 0 et 10 FA/mot/heure.

tions, fourniture vocale de la clé de recherche. Ce dernier point est le plus novateur et va permettre une meilleure ergonomie des outils de recherche d'informations.

La recherche de mots clés s'effectue en 2 étapes. La première étape (off-line) consiste à indexer préalablement la base de données par des phonèmes en exploitant les techniques markoviennes usuelles de reconnaissance de la parole développées au CNET. (On a utilisé un modèle syllabique de reconnaissance phonétique). La deuxième étape (on-line) compare les séquences phonétiques de la base de données avec celle de la requête prononcée. Les deux algorithmes développés dans la suite de cet article traitent uniquement cette phase de recherche rapide. L'indépendance au locuteur est englobée dans la phase *off-line*, tandis que l'indépendance au vocabulaire est inhérente au décodage phonétique à grammaire nulle du corpus.

Il convient de souligner que cette méthode de recherche *on-line* à partir d'index phonétiques procure une meilleure homogénéité des scores que lors de l'utilisation d'un HMM-spotter, car les différences non pertinentes sont absorbées par l'étage phonétique. En effet, deux [a] prononcés différemment ont des scores acoustiques différents, mais une fois classifiés [a], ils sont indiscernables pour la suite des calculs. On favorise ainsi la discrimination d'un mot clé du flux de parole englobant. Rappelons que le phonème est l'unité permettant de distinguer deux mots différents.

Pour bien cerner la difficulté de détection d'un mot clé fourni oralement, rappelons que les meilleurs décodeurs phonétiques actuels ont des taux d'erreur qui avoisinent 50 %. On peut donc imaginer une transcription phonétique de la base avec un phonème sur deux erroné, et une requête vocale subissant la même dégradation. Cette symétrie des erreurs possède tout de même un aspect positif : la suppression des erreurs systématiques comme la confusion régulière d'un phonème avec un autre. Nous allons ainsi construire dynamiquement un modèle de Markov à partir du décodage phonétique du mot à rechercher.

3. MODÈLE DE BASE — ALGORITHME 1

3.1 Principe

L'algorithme 1 consiste à comparer en mode spotting[2] deux séquences de phonèmes et à en établir une distance pour juger de leur dissemblance. Si la distance est suffisamment faible (en deçà d'un certain seuil), on déduira la détection du mot clé.

Pratiquement, les enregistrements indexés de la base de données (suites de phonèmes) sont décodés par le modèle de Markov du mot clé (Action d'alignement). Ce dernier est construit dynamiquement à partir des phonèmes issus de la reconnaissance automatique (qui comportent des erreurs), et se compose d'un enchaînement de mini-modèles du type suivant :

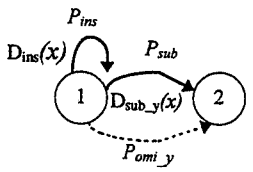


Figure 1: Modèle phonétique élémentaire du phonème *y*

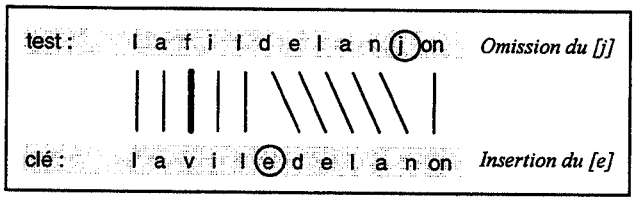


Figure 2: Alignement possible pour « la ville de Lannion »

Lors de l'alignement des séquences de phonèmes on va considérer que chacun a la possibilité d'être substitué avec un autre, ou simplement d'être omis ou inséré (figure 2). On définit ainsi des probabilités d'omission, d'insertion et de confusion de phonèmes, regroupées dans une matrice de confusion.

Les densités de probabilité discrètes *Dins* et $Dsub_y$ sont issues de l'apprentissage de la matrice de confusion sur une base de données. Les probabilités de transition proviennent du même apprentissage. Le score obtenu en sortie du modèle est normalisé par le nombre de transitions empruntées lors de l'alignement.

Apprentissage des paramètres Ne disposant pas de base de données française spécifique pour faire de la détection de mots clés, nous avons utilisé le corpus d'expressions *all* (CNET) issu d'enregistrements téléphoniques, et comportant 684 phrases courtes prononcées par différents locuteurs, soit 29254 enregistrements au total. Cette base phonétiquement équilibrée a permis de créer des fichiers d'alignements corrects et incorrects de mots isolés.

Nous avons testé différentes méthodes d'apprentissage. Tout d'abord la maximisation de la vraisemblance des observations avec l'algorithme EM (Expectation Maximisation), puis ensuite l'optimisation des paramètres par algorithme génétique[3] de sorte à accroître la discrimination des alignements corrects et incorrects. Nous avons ainsi optimisé pour les phonèmes de la langue française, 1683 paramètres.

3.2 Symétrisation analytique des paramètres

Dans le cas où on ne dispose pas des enregistrements sonores des phrases et des mots clés, l'apprentissage de la matrice de confusion peut être réalisé à partir d'une base de données phonétiquement étiquetée classique. Nous

[2] c'est-à-dire entre deux boucles de rejet

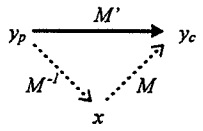
[3] optimisation génétique faite sur une population de 10 individus de 1683 paramètres, avec un taux de mutation de 0,5 et un taux de crossover de 0,3. On peut assimiler cet apprentissage à une technique de recuit simulé à têtes multiples + crossover.

proposons dans ce paragraphe une méthode analytique pour transformer la matrice dissymétrique résultante en matrice symétrique.

Soit x le processus de transcription phonétique exacte des mots. En émettant l'hypothèse d'indépendance conditionnelle en x, de y_c et y_p les processus d'émission des phonèmes des mots clés et des phrases, on peut écrire :

$$P(y_c|y_p) = \sum_x P(y_c|x).P(x|y_p) = \sum_x P(y_c|x)\frac{P(y_p|x).P(x)}{P(y_p)}$$

D'un point de vue plus pratique, on obtient les paramètres du modèle symétrique M' à partir des estimateurs des matrices dissymétriques $y|x$ (on retourne le modèle $y_p|x$) :



On calcule les probabilités d'insertion d'un phonème comme suit :

$$\forall i, Pins_{M'}(i) = \underset{y_c|x}{Pins}(i) + \sum_k \underset{y_c|x}{Psub_k}(i).\underset{x|y_p}{Pins}(k)$$

Les calculs de $Pomi_{M'}(j)$ et de $Psub_{M'}(i)$ sont faits de façon similaire :

$$\forall j, Pomi_{M'}(j) = \underset{x|y_p}{Pomi}(j) + \sum_k \underset{x|y_p}{Psub_j}(k).\underset{y_c|x}{Pomi}(k)$$

$$\forall ij, Psub_{M'j}(i) = \sum_k \underset{y_c|x}{Psub_k}(i).\underset{x|y_p}{Psub_j}(k) + \underset{y_c|x}{Pins}(i).\underset{x|y_p}{Pomi}(j) + \sum_k \underset{y_c|x}{Pomi}(k).\underset{x|y_p}{Pins}(k)$$

3.3 Résultats

La base de test est consituée d'alignements corrects et incorrects (c'est-à-dire qu'elle met en correspondance les chaînes phonétiques de mots identiques et différents, respectivement). On souhaite la plus forte discrimination possible entre les deux classes. On mesure ainsi le taux d'erreur de classification. L'évaluation des résultats se fait à partir d'un seuil de classification (sur le score normalisé) qui minimise le taux d'erreur sur la base d'apprentissage. Tous les alignements dont le score est inférieur à ce seuil sont considérés comme des cas de détection du mot clé recherché.

Table 1: Résultats de l'algorithme 1 sur la base *all*

taux d'erreur (en %)	***all***
matrice unité (DP-Matching)	12,33
maximum de vraisemblance (MV)	8,21
symétrisation analytique	7,37
optimisation génétique	6,4

Ce tableau met bien en évidence le gain obtenu avec un apprentissage discriminant par rapport au maximum de vraisemblance (22 %) ou au DP-matching (48 %).

On observe également une détection plus robuste avec la méthode de symétrisation analytique (7,37 % au lieu de 8,21 %). En effet, les paramètres étant appris sur des alignements comportant une phonétisation exacte, une des deux sources d'erreurs est supprimée. On obtient de surcroît un taux d'erreur relativement bas à l'apprentissage au maximum de vraisemblance de la matrice dissymétrique : **2,56 %**. Ce chiffre représente les performances du programme dans le cadre d'une utilisation comme système de reconnaissance à vocabulaire illimité, ce qui paraît plutôt prometteur pour la reconnaissance à deux étages.

4. Modèle amélioré — Algorithme 2

Ce nouvel algorithme prétend fournir une plus forte robustesse des alignements en se rapprochant de la façon humaine de les réaliser.

4.1 Principe

On va pouvoir faire correspondre plusieurs phonèmes consécutifs d'une chaîne à un phonème de l'autre chaîne.

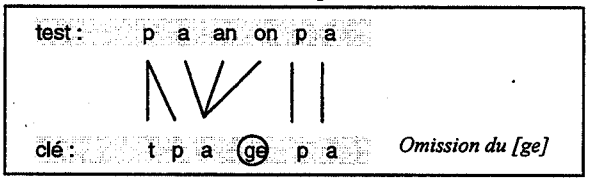
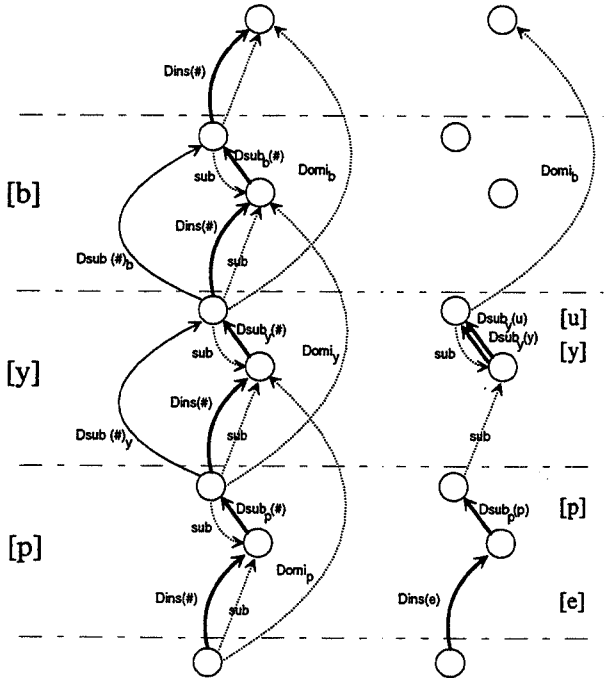


Figure 3: Correspondances de phonèmes possibles pour le mot « papa »

Pour cela, on utilise une topologie légèrement plus complexe (introduction d'un nouveau type de transition) mais sans augmenter le nombre de paramètres. Par souci de clarté, les probabilités de transition sont non représentées (figure 4).



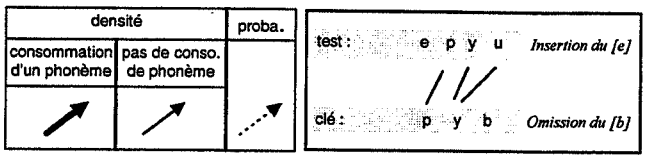


Figure 4: Modèle symétrique du mot « pub », et exemple d'alignement avec les phonèmes [epyu]

Apprentissage Nous avons effectué trois types d'apprentissage sur ce modèle : maximum de vraisemblance, correctif (méthode hybride dérivée de [Bah88]) et optimisation génétique.

4.2 Résultats

Table 2: Résultats de l'algorithme 2 sur la base *all*

taux d'erreur (en %)	***all***
maximum de vraisemblance (MV)	9,59
apprentissage correctif	6,27
optimisation génétique	6,08

La méthode corrective (discriminante) nous fait chuter de 35 % par rapport au MV. Même si le taux d'erreur est proche de celui de l'algorithme 1 (6,4 %), les alignements sont de meilleure qualité dans l'ensemble.

5. ÉVALUATION EN MODE *SPOTTING* SUR AGS

Pour tester les résultats en mode spotting (c'est-à-dire sur des alignements *mot-phrase*), nous avons créé une autre base de données à partir des enregistrements de la base AGS. C'est un corpus de parole spontanée recueillie par téléphone lors du test d'un système de dialogue. Cette base contient parfois des informations perturbées (bruits de bouche, portes, autres personnes présentes dans la pièce) et les voix sont quelquefois faibles ou criées. Cette base a tout de même l'avantage de représenter des conditions réelles d'exploitation. Nous avons sélectionné 1199 phrases longues et 53 mots clés prononcés isolément.

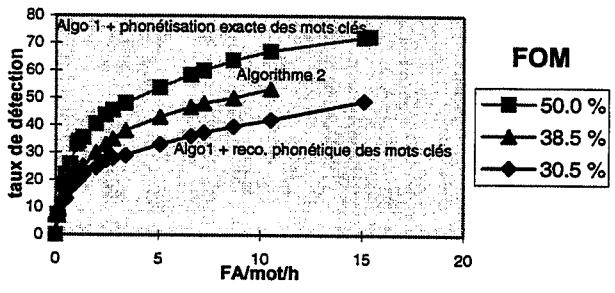


Figure 5: FOM des deux algorithmes sur AGS

En mode spotting, les résultats sont moins probants. En effet, l'absence de contrainte sur les positions de début et de fin augmente la probabilité de trouver un segment approchant ; le taux de fausses alarmes croît rapidemment.

On note que lorsque les mots clés ne sont plus issus de la reconnaissance phonétique, le FOM est de 50 %. On se rapproche alors du système de [Jam94] ou de [Jun96]. Étant donné la difficulté de notre base AGS, les résultats sont jugés assez satisfaisants.

6. CONCLUSION ET PERSPECTIVES

Notre détecteur de mots vocaux se base sur une reconnaissance à deux étages : un premier étage qui indexe *off-line* tous les enregistrements de la base, et un deuxième étage qui recherche *on-line* un mot clé à partir d'une chaîne de phonèmes issue de la reconnaissance phonétique. Cette méthode permet la mise en place instantanée de systèmes de reconnaissance de la parole sans apprentissage ni compilation de modèles, ou encore l'extraction rapide de documents sonores interactive avec l'utilisateur. En effet, la vitesse de recherche des informations est environ **8000 fois** plus rapide que le temps-réel sur une machine Sun UltraSparc à 200 Mhz. On note également l'absence de modèle poubelle pour la parole de fond, ce qui rend le système totalement indépendant du signal englobant le mot clé.

Les perspectives d'évolution du système sont multiples. Comme on souhaite s'affranchir totalement du vocabulaire, il est difficile d'améliorer le décodage phonétique avec des contraintes lexicales. Cependant la pondération du score de certains couples de phonèmes alignés pourraît augmenter la discrimination des alignements de type fausse-alarme. On pourra par exemple accorder plus d'importance à l'alignement des voyelles. Sans omettre les techniques de rescoring au niveau signal des tronçons détectés de sorte à éliminer les fausses alarmes, on peut mettre en place des techniques de clusterisation de phonèmes, ou même calculer un jeu d'unités sub-phonétiques optimal en matière de détection de mots. Ce travail se poursuit actuellement et vise une approche indépendante de la langue.

BIBLIOGRAPHIE

[Bah88] L. R. BAHL, P. F. BROWN, P. V. DE SOUZA, R. L. MERCER, « A New Algorithm for the Estimation of Hidden Markov Model Parameters », I.C.A.S.S.P. 1988 - pp. 493-496

[Gel96] P. GELIN, C. J. WELLEKENS, « Keyword spotting enhancement for video soundtrack indexing », I.C.S.L.P. 1996

[Hig85] A. L. HIGGINS, R. E. WOHLFORD, « Keyword recognition using template concatenation », I.C.A.S.S.P. 1985 - pp. 1233 à 1236

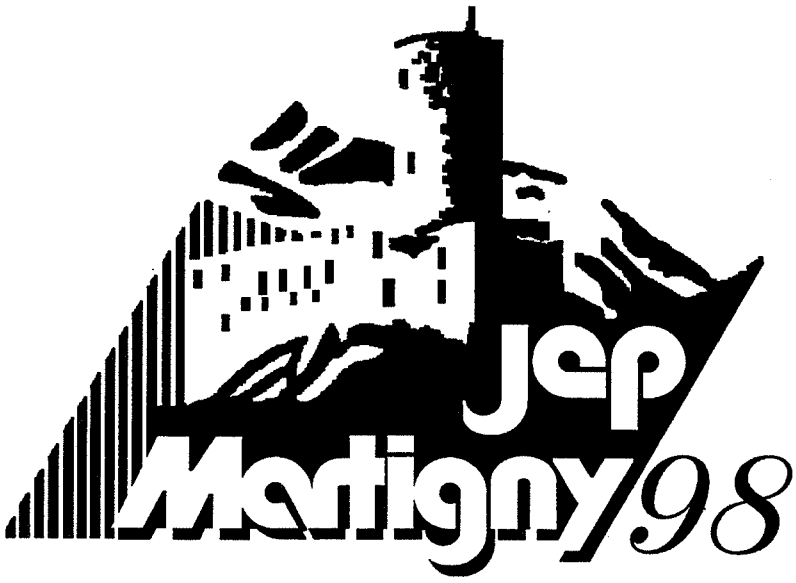
[Jam94] D. A. JAMES, S. J. YOUNG, « A fast lattice-based approach to vocabulary independant wordspotting », I.C.A.S.S.P. 1994 - pp. 377 à 380

[Jun96] J. JUNKAWITSCH, L. NEUBAUER, H. HÖGE, G. RUSKE « A new keyword spotting algorithm with pre-calculated thresholds », I.C.S.L.P. 1996

[Ros90] R. C. ROSE, D. B. PAUL, « A Hidden Markov Model based keyword recognition system », I.C.A.S.S.P. 1990 - pp. 129 à 132

[You96] S. J. YOUNG, « Audio Document Indexing and Retrieval », Cambridge University Engineering Department - août 1996

Conférence invitée



Spoken Language Dialogue : Architectures and Algorithms

Roberto Pieraccini, Esther Levin, Wieland Eckert

AT&T Labs-Research
180 Park Avenue, Florham Park, NJ 07932, U.S.A
Tél.: (973) 360-8563 - Fax: (973) 360-8092
e-mail: {roberto, esther, eckert}@research.att.com - http://www.research.att.com

ABSTRACT

In this paper we discuss the evolution of spoken language dialogue architectures and algorithms at AT&T Labs-Research. First we introduce a user initiated system that is able to understand user requests and provide the correct answers in context. Then we discuss a mixed initiative system that can change the course of the interaction by asking clarification questions in order to reach the application goal in the most effective way. Finally we introduce a mathematical formalization of human./machine dialogue systems that can lead to one of the most ambitious goals in the field today, namely learning the dialogue strategy from interactions.

1. INTRODUCTION

One of the ultimate goals of the speech recognition community is the implementation of machines able to interact with humans in a natural conversational way in order to provide services that would otherwise require human operators or unfriendly menu based systems.

Although research in the field of human-machine spoken dialogue started decades ago, only recently [9][15][19] were we able to implement end-to-end systems. The ability of building such systems can be attributed without any doubts to the recent advances in the related technologies. First, the ability of building high accuracy large vocabulary spontaneous speech recognizers that can work in real time on inexpensive computers. Second, the development of robust language understanding systems that can deal with the problems inherent to spontaneous speech. Third, the recent flourishing of research aimed at the problem of dialogue system architectures, and trying to solve issues such as modularization, porting, reusing, fast prototyping, etc.

In this paper we will discuss some of the evolution of dialogue systems research at AT&T Labs-Research. This effort started with the implementation of language understanding systems suited for spontaneous speech that could be regarded as a primitive form of dialogue system that allow only user-initiated interactions. Then a more complex architecture for mixed-initiative dialogue was investigated and implemented. Finally we proposed a mathematical formalization of the human-machine dialogue that will allow for learning of dialogue strategies.

2. LANGUAGE UNDERSTANDING AND USER-INITIATED DIALOGUE

A user-initiated dialogue system consists of a machine that responds to user's questions with the best possible answer given the context of the whole session. The evolution of dialogue is then a succession of user's questions and system's answers. User initiated dialogue can be generally implemented through a cascade architecture, an example of which (for a database retrieval application) is shown in Figure 1.

2.1 Speech Recognizer

We will not describe the speech recognizer in this context, except saying that is a HMM sub-word based recognizer [6] using a variable n n-gram stochastic language model [7] that produces the first best sentence hypothesis in the for of a plain string of text.

2.2 Lexical Analyzer

The function of the lexical analyzer is that of creating a lattice of word hypotheses out of a string of words (i.e. the transcription of speech into words generated by the speech recognizer [5].

The different hypotheses in the lattice correspond to possible interpretations of words and/or phrases according to predefined semantic categories. Most of the categories correspond to classes of attributes of the database (e.g. cities, dates, etc.) words that we decided to consider as synonyms for later stages of the process. For instance, the distinction between singular and plural forms of words as well as different verb inflections is irrelevant for many tasks, so they are collapsed in the same category. This both reduces the effective size of the lexicon and enhances the robustness of the stochastic conceptual representation [1], especially in presence of recognition errors (speech recognizers are likely to confuse among different inflections of the same word). Numbers and acronyms are parsed in all the possible ways, each parse constituting a separate lattice hypothesis.

2.3 Conceptual Decoder

The idea behind the conceptual decoder is that of treating a sentence as a sequence of phrases corresponding to units of meaning called concepts. A Markovian process models the sequence of concepts (a first order process in the current implementation).

The sequence of words that forms a phrase related to a given concept is also modeled by a Markovian process represented by a concept conditional n-gram language model.

In the current implementation the concept conditional language models are bi-gram back-off networks [6]. The overall model is called conceptual HMM [1][3].

The function of the conceptual decoder is that of segmenting a sentence into phrases and assigning each

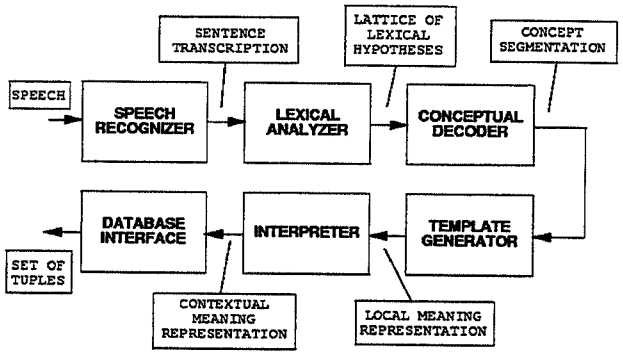


Figure 1: The CHRONUS spoken language understanding system.

phrase to the correspondent concept. This translates into finding the most likely sequence of states in the conceptual HMM given the lattice produced by the lexical analyzer and it is implemented as a finite state automata intersection operation as explained in [4].

There are two problems in designing a conceptual decoder for a given application: the choice of the conceptual units and the training of the conceptual HMM. The choice of the conceptual units generally requires a good knowledge of the task. Most of the units generally correspond to entities of the database. Training a conceptual HMM implies providing a considerable amount of examples of conceptual segmentations and running a Viterbi training algorithm [1] in order to estimate the parameters of the model. Of course the most difficult part is providing examples of segmentations, namely segmenting by hand thousands of sentences. This can be greatly alleviated using a training loop procedure [3] that uses an automatic evaluation criterion to detect possibly wrong segmentations in a segmented corpus.

2.4 Template Generator

The template generator produces a representation of the sentence semantics in the form of a template, i.e. an unordered set of keyword/value pairs (called *tokens*) [3] based on the segmentation produced by the conceptual decoder.

Most tokens correspond to attributes of the database and their values. For instance the token DEPARTURE_CITY:SSFO assignes the value SSFO (i.e symbol for the city of San Francisco) to the database attribute *departure_city*.

There are tokens that represent modifiers and are used either by the interpreter or by the database interface. Examples of this are the token QUESTION (whose value YES/NO instructs the database interface to produce a yes/no answer), or the token DEPARTURE_PERIOD (whose value modifies any DEPARTURE_TIME token in the same template to AM or PM accordingly).

The information to be extracted from the database is specified by the keyword SUBJECT.

2.5 Interpreter

The function of the interpreter consists in resolving ambiguities and providing missing information in the current template.

In general the interpreter merges the current template with a *context* template according to a set of masking rules. The resulting template is used for further processing and

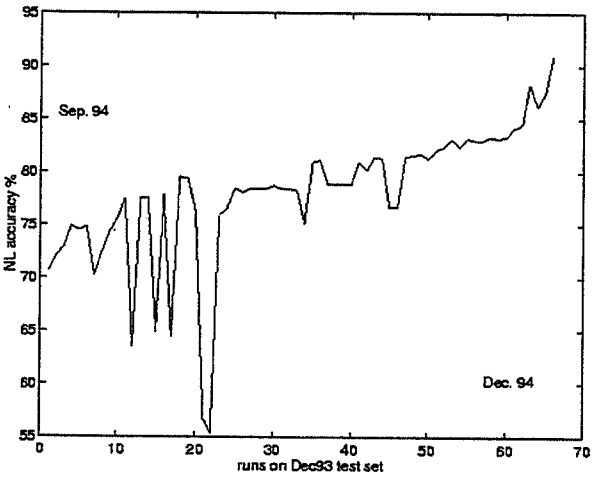


Figure 2: Incremental improvement of the ATIS system .

stored as the new *context* template.

One example of *general* masking rule states that tokens in the most recent template mask tokens in the context template with the same key. One example of a *specific* masking rule states that a YEAR token in the most recent template masks DATE tokens in the context template (i.e. the specification of a different year makes the previous date invalid).

The interpretation rules are (in our implementation) written by hand and special care should be taken for them to be consistent. A corpus and an automatic evaluation (such as the one that was provided in the ARPA ATIS project [2]) are very useful for developing of this module, since they allow to set up an automatic way of assessing the effectiveness and the consistency of each newly introduced rule. In the development of the ATIS interpreter, every time new rules were introduced the system was evaluated on a corpus of about 5,000 sentences (this test

was completely automated and required only a few hours of CPU time).

Figure 2 shows the percentage of correct sentences in almost 70 experimental runs during the three months in which the ATIS interpreter was developed.

2.6 Database Interface

The database interface takes the semantic representation in the form of a template and extracts the requested information from a relational database. In general it is necessary to provide some specific piece of software that will convert the template representation into the query language of the database (e.g. SQL, web-based forms, etc.). Other more general solutions can be adopted in the case the data is local and represented in a relational form [13].

1.1 System Performance

The performance of the system based on the CHRONUS approach is shown in Figure 3 where the official figures for the three standard tasks in the ATIS program are reported [5]. The scores for the speech recognition task, the text understanding task and the speech understanding task were among the best across all sites participating to the evaluation.

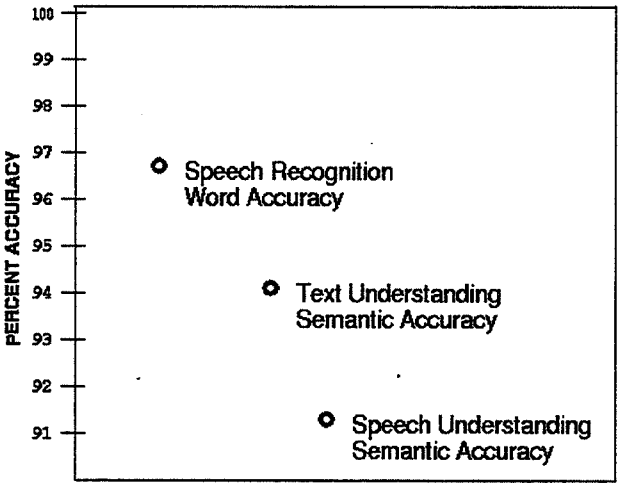


Figure 3: Performance of the AT&T system in the 1994 ARPA ATIS official evaluation.

3. Mixed Initiative Dialogue Architecture

A user-initiated system has many obvious limitations for complex applications like database retrieval. For instance when an initial question like *I need some information about flights to New York* is asked, a system that does not have the possibility of asking questions (such as a user initiated system) will respond showing all the flights to New York from any other city in the database. Such a long response may not be appropriate for the kind of output channel in use (e.g. an audio response via TTS). Moreover the cost of accessing a remote database with such an unrestricted question and the time for transferring the results can be excessive.

Mixed initiative systems constitute the natural evolution of user-initiated systems. A mixed initiative system can ask questions as well as provide answers, and can change the course of the conversation when it is necessary. Table 1 shows an example of a mixed initiative interaction.

Mixed initiative can hardly be implemented by simple cascade architectures. Moreover it is desirable to have a *general* architecture that would easily fit different applications, and would allow for a rapid development of new systems.

AMICA [13], the AT&T Mixed Initiative Conversational Architecture, is based on the identification of a set of general, logically motivated functions, called *dialogue actions*. We think that for certain classes of dialogues there is a finite (and small) number of such actions that can be implemented in a general way and parameterized in order to be used for different applications.

The following is a basic set of dialogue actions:

- **Understand**: transforms the user natural language input into a semantic representation (it corresponds to the speech recognizer, lexical analyzer, conceptual decoder, and template generator described in the previous section).
- **Verbalize**: transforms a semantic representation into natural language sentences. The action of verbalizing data retrieved from the database is called **verbalize data**.
- **Context Tracking**: consists in the interpretation of the current user input in terms of the history of the interaction. It includes both *dialogue interpretation*, namely the ability of dealing with expectations set by the machine itself (e.g. disambiguation on the basis of previous questions asked by the machine), and *discourse interpretation*, namely the ability of taking into account the context set by the user during the course of the whole interaction (corresponding to the *interpreter* described in the previous section). In both cases it is necessary to be able to recognize ambiguities and ask the proper disambiguating questions.
- **Retrieve Data**: consists in the assembly and submission of a query to a local or remote database according to the current information gathered by the system.
- **Constrain:** generating a request for additional constraints based on the current topic and the available.
- **Relax:** relaxing attributes of the query in order to retrieve approximate answers from the database.

3.1 Dialogue Control

For explaining how the AMICA architecture works it is necessary to introduce the concept of system *state*. The system state is the collection of all the pieces of informa-

tion that characterize the whole dialogue system at a certain point in time. The system state univocally determines the continuation of the interaction.

The system state is represented in AMICA by a frame like (i.e. key/value pairs) data structure, and example of which is shown in Figure 4.

The *dialogue strategy* selects one of the possible actions based on the current state of the system,. As an effect of the action, the system state changes, then a new action is selected by the strategy, etc.

In general, in a realistic application, it is not possible to enumerate of all the states and assign each one of them to a particular action. Therefore the selection of the proper action given the current state of the system is based on conditions on the current state. In AMICA the strategy is implemented as a finite state recursive transition network whose nodes are associated with actions and whose (ordered) transitions are correspond to conditions on the current state.

```
USER: {
  CITY: NEWARK
}
CONTEXT: {
  SUBJECT: GROUND TRANSPORT
  ORIGIN_CITY: DENVER
  DESTINATION_CITY: NEWARK
  TIME_PERIOD: MORNING
  AIRLINE: DELTA
}
EXPECTATION: {
  GROUND_CITY: ?
}
QUERY: {
  SUBJECT: GROUND TRANSPORT
  GROUND_CITY: NEWARK
}
DATA: {
  TRANSPORT: TAXI
  TRANSPORT: LIMOUSINE
  TRANSPORT: RENTAL CAR
}
```

Figure 4: Frame representation of the dialogue state

An example of simple strategy for mixed initiative dialogue is shown in Fig. 5 (*Context Tracking* has been included in the *Understand* action in this example).

The strategy starts with the *verbalization* of a greeting. *Understand* gets user's input (that is interpreted by the *context tracking* function). If the current user request is under-specified, *Constrain (1)* will generate a request for additional constraints that will be appropriately verbalized. Otherwise *Retrieve* will access the database with the current request. If too much data is returned by the database (*NDATA* > *t*), then *Constrain (2)* will generate a request for additional constraints that will be verbalized. If the database returns no data (NDATA == 0), *Relax* will attempt to relax constraints until some data is retrieved. If a reasonable amount of data is retrieved (0< NDATA <= t), it will be verbalized and a new request from the user will be collected.

This simple strategy is able to generate a behavior that evolves through three interaction modes, corresponding to loops in the strategy of Figure 5. The first loop, that we call *minimal information* (Understand – Constrain (1) – Verbalize) will gather enough information from the user in order to perform a reasonable database query. In fact, the absence of this behavior will force to retrieve data even when the user did not give the minimal required information (e.g. the user asking "I would like information about flights to New York", and the system requesting from the database *all the flights* to New York from any airport, a useless expensive query!).

The second loop, that we call *constraining*, (Understand – Constrain (1) - Retrieve – Constrain (2) - Verbalize) will attempt to reduce the amount of data *presented* to the user. Depending on the channel, there is a maximum amount of data the user will be able to assimilate. Typically two or three pieces of information (tuples) with an audio channel, 10 or more with a small display, etc. If more data is retrieved, other questions need to be asked in order to trim it down to a reasonable quantity.

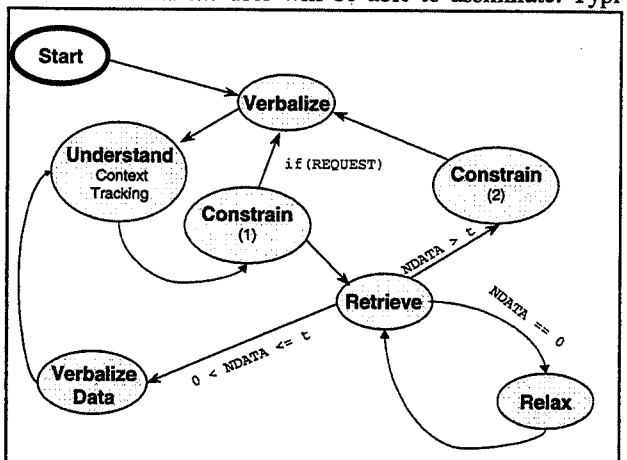


Figure 5: Example of mixed initiative strategy

Finally the third loop, the *relaxation* loop (Understand - Constrain (1) - Retrieve – Relax – Retrieve .... Relax – Retrieve – Verbalize Data) attempts to give the user some data in presence of over-constrained request (e.g. *I would a flight that leaves (exactly) at 7 a.m.*).

An example of a fragment of dialogue drawn from the Airline Information System and using the strategy of Figure 5 is shown in Table 1.

3.2 Portability Issues

The process of developing a dialogue system for a new application is greatly eased by the functional independence of the various modules that compose the system architecture. Moreover all the actions described so far can be implemented in a *table driven* or *programmable* fashion. We will give here some examples of the action programmability in order to stress the application independence of the system.

3.2.1 Verbalize

The verbalization (or sentence generation) process is implemented by an interpreter that accepts sentence archetypes with conditional clauses such as:

```
if(NDATA > 1) {"There are "}
       else {"There is "} NDATA
if(NDATA > 1) { "flight " } else {"flights"}
  "from " DEPARTURE_AIRPORT
        "to " ARRIVAL_AIRPORT ".";
```

This archetype can generate sentences like:

There are 5 flights from Newark to Denver.
There is 1 flight from San Diego to Boston.

For the verbalization of relational database tuples the process relies on the semantic description of the database augmented with linguistic information. For instance verbs and prepositions are attached to pairs of database attributes, like:

```
FLIGHT_ID  flies to DEPARTURE_AIRPORT
FLIGHT_ID leaves from ARRIVAL_AIRPORT
FLIGHT_ID is served by AIRLINE
FLIGHT_ID uses a AIRCRAFT
```

These predicates are organized into a finite state network that is then used for building sentences with an arbitrary number of attributes, like

Flight UA706 leaves from Denver, flies to Newark, is served by United, and uses a Boeing 747.

3.2.2 Constrain

The constrain action, in the way it was implemented in this system, requires the specification of logical expression whose truth guarantees that a set of constraints for a given topic is available in the current state. For instance, the following is a valid set of constraints for the topic FLIGHT:

```
if(SUBJECT: FLIGHT) askFor
  (AIRLINE && FLIGHT_NUMBER) ||
(DEPARTURE_AIRPORT &&
          ARRIVAL_AIRPORT);
```

The constrain algorithm matches this expression with the current state (each symbol is true if the corresponding attribute is present in the state). If the expression is false, another logical expression is computed whose truth will make true the original expression. For instance, if the state includes the following information:

```
AIRLINE : DL
DEPARTURE_AIRPORT: EWR
```

The constrain algorithm will generate the following expression:

```
MISSING: FLIGHT_NUMBER || ARRIVAL_AIRPORT
```

Corresponding to the question:

Please specify your flight number or arrival airport

3.2.3 Relax

The relaxation action, in its simpler implementation, requires a list of domain attributes ordered according to their relative importance. For instance, in the airline information task, the list can be: origin/destination, date, time, type of aircraft, and meal. Attributes are relaxed in

1	VERBALIZE	This is the AT&T Chronus Air Travel Information System, how may I help you?
2	UNDERSTAND	**I want a direct flight from Denver to Boston between eight and nine thirty tomorrow morning**
3	RETRIEVE RELAX RETRIEVE VERBALIZE DATA	There are no flights from Denver to Boston leaving between 8 and 9:30 in the morning. Instead there are flights leaving at 10:39 and 11:05. Do you have any preference?
4	UNDERSTAND	**Give me a delta flight**
5	UNDERSTAND RETRIEVE VERBALIZE	Flight DL296 leaves at 10:39 am and arrives at 3:15 pm. Other questions?
6	UNDERSTAND	**I need ground transportation**
7	CONSTRAIN VERBALIZE	Where, Denver or Boston?
8	UNDERSTAND	**Boston**
9	RETRIEVE VERBALIZE DATA	The following ground transportation is available in Boston: taxi, limousine, rental car.
10	UNDERSTAND	**Please list all the flights with Continental**
11	RETRIEVE CONSTRAIN VERBALIZE	There are too many flights that match your request. I need more information. At what time you want to leave?

Table 1: Example of dialogue with the AT&T Arline Information System

the order until a non empty set of tuples is found.

In certain application the relaxation procedure is more complex, and special ad-hoc functions have to be designed. For instance, in a movie locator application, one can be interested in relaxing the *city* attribute, hence finding which one is the theater *closer* to a given city that shows a certain film. This requires a specific algorithm that computes the distance of theaters from cities and searches for the closest one.

4. Formalization and Learning of Mixed Initiative Dialogue Systems

As we saw in the previous section the strategy determines the behavior of the system, and it is generally built by hand based on a good knowledge of the domain. For instance a reasonable strategy for the airline domain is to request the origin and the destination of the flight and the departure date at the beginning of the transaction, because we know that those are the most distinctive questions that will greatly restrict the result of the database search. But when faced with a new domain the choice cannot be that obvious. Moreover the criterion on which we base our reasonable strategy (e.g. length of the interaction and database retrieval in the airline example) can be different from task to task. It would be highly desirable to have a mathematical formalization of dialogue system, and to be able to derive objective criteria for evaluating a strategy, and eventually learn the optimal one.

4.1 Dialogue Systems as Markov Decision Processes – a Tutorial Example

We will use a simple tutorial example for explaining the formalization that we introduced [14] for describing dialogue systems. We will assume that the goal of our tutorial dialogue system is getting a date from a user through the *shortest* possible interaction.

We will formalize this dialogue system by describing it in terms of a **state space**, **action set**, and **strategy**.

In our simple tutorial example, the **state** of the system includes only two entries: the day and the month, whose values can be either empty, or filled through interaction with the user. The total number of states is 411, including one empty initial state, 12 states for which the month is filled and the day is not, 31 states in which the day is filled, but not the month, 366 states with complete dates, and a special final empty state.

The action set include only four possible actions:

1. Asking **which day**.
2. Asking **which month**.
3. Asking **which date**. This is a more open-ended question and will compel the user to answer with a day and a month in the same sentence.
4. A **final action**, closing the dialogue and submitting the information.

In actions 1, 2 and 3, the system asks the question, and then activates a speech recognition system in order to obtain the user's answer. We assume that the probability of error of the speech recognizer depends on the kind of question. In particular questions 1 and 2 will both show a different word error probability than question 3 (being the user's answer longer and more articulated). In particular we will assume that a question about a single attribute (i.e. action 1 or 2) will have a word error probability p_1, while a composite question (i.e. action 3) will have a word error probability p_2, with $p_1 < p_2$.

Once the system is in a given state (e.g. the initial empty state) and an action is executed (e.g. the system asks *which month*), the system goes into another state (e.g., assuming that the user is compliant the system will go to one of 12 possible states depending on the user's answer). Since the new system state resulting from the execution of an action depends on external causes (i.e. the user's response and the recognizer accuracy), we model it by transition probabilities $P_T(s_{t+1} = s' \mid s_t = s, a_t = a)$.

A **dialogue session** corresponds to a path in the state space starting at the initial state and ending at a final state. The **dialogue strategy** specifies, for each state reached, what is the next action to be invoked.

At this point we introduce an objective measurement of the quality of a dialogue session based on a cost function:

$$C = \sum C_i, \quad (1)$$

i.e. the sum of individual costs incurring during the interaction. The individual costs *Ci* measure, for instance, the effectiveness of the interaction (e.g. its duration), the distance from the application goal (e.g. whether the system provided the user the requested service), and other objective costs, like database access costs. Furthermore it has been shown [13] that also subjective measurements like

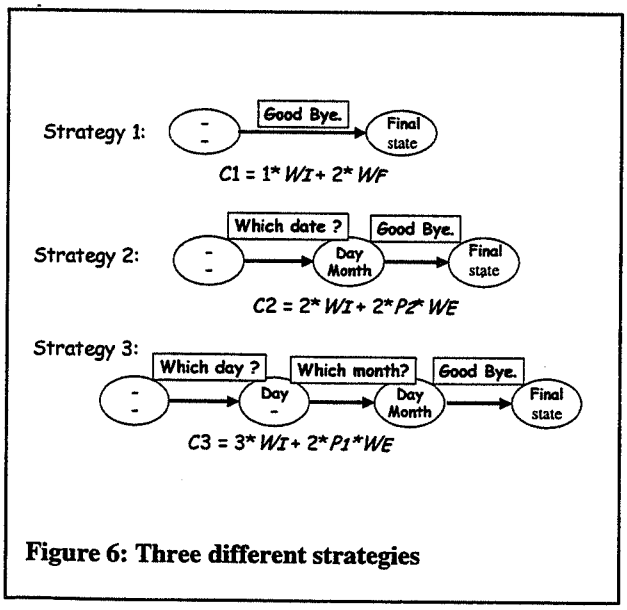


Figure 6: Three different strategies

user satisfaction can be modeled as a linear combination of costs, and could therefore used as one of the terms in equation (1).

Now we can define that an optimal dialogue strategy is the one that minimizes the expected value of cost *C*.

For our tutorial example, where the goal of the system is to obtain the correct day and month values through the shortest possible interactions, the objective function includes three terms:

$$<C>=W_I \cdot <N> + W_E \cdot <E> + W_{UF} \cdot <UF>$$

Where <*N*> is the expected duration (i.e. expected number of interactions) of the dialogue; <*E*> is the expected number of errors for the filled slots (ranging from 0 to 2); and <*UF*> is the expected number if unfilled slots. The optimal strategy, i.e. the one that minimizes <*C*>, will be a trade off between duration and errors (in fact asking question 3 rather than 1 and 2 will produce a shorter but more error prone dialogue), and will depend on the actual values of the error probabilities.

In order to reflect this objective function in our dialogue model, we associate a cost *c* to action *a* selected while in a state *s* (in general we can say that action costs are described by conditional distributions:

$$P_C(c_t = c \mid s_t = s, a_t = a)).$$

The cost incurred with any of the first three actions in day-and–month dialogue system is $W_I + W_E \cdot E$. If we assume that the concept error rate for recognition of month or day for questions 1 and 2 is p_1, for question 3 is p_2, with $p_2 > p_1$, then the expected cost accumulated when actions 1 or 2 are taken is $W_I + W_E \cdot p_1$, while it is $W_I + 2 \cdot W_E \cdot p_2$ for action 3. The cost of action 4 (closing the dialogue and submitting the obtained date) depends deterministically on the state in which this action is taken and is $W_I + 2 \cdot W_{UF}$ for an initial state, $W_I + W_{UF}$ for states in which either the day or month is unfilled, and W_I for states in which both values are filled.

Of course, different strategies for the same system result in different expected session costs. Figure 5 shows three different strategies and their associated expected costs. For example strategy 1 (where the system does not even engage in dialogue, closing the dialogue as the first action) is optimal when the recognition error rate is too high, i.e. $p_1 > (W_{UF} - W_I)/W_E$.

In strategy 2 the system opens the dialogue by asking the open ended question 3, it fills out the day and the month slots with the values recognized from the user response, and then closes the session. In strategy 3, the system first fills out the day and then the month by engaging in actions 1 and 2, and then closes the session. Strategy 3 is optimal when the difference in error rates justifies a longer interaction: i.e. when $p_2 - p_1 > W_I/(2 \cdot W_E)$.

The model we just described is known as Markov decision process[8], or MDP, and is in general described by the quadruple: state space, action set, transition probabilities, and cost distributions.

4.2 Automatic Design of Man-Machine Dialogue Systems: the Reinforcement Learning Approach.

Stating the problem of man-machine dialogue design as an optimization problem provides the following potential advantages:

Objective evaluation: It is possible now to assess different strategies for the same system just by comparing their expected costs. It is also possible to compare different systems that share the same objective function.

Automatic design: Since the problem of strategy design is cast as an optimization problem, it is possible to devise methods for performing this optimization automatically. In the simple example of the previous section it was possible to derive the optimal strategy analytically. This is not true in general with realistic applications. However there are learning techniques, such as reinforcement learning [8], that allow for finding the optimal strategy in an MDP from interactions.

4.2.1 An experiment with reinforcement learning.

Rather than explaining here the details of the reinforcement learning algorithm, we will discuss the results of an experiment devoted at automatically finding a strategy for a simplified version of the airline information task.

We restricted the possible actions of the system to the following:

Greet the user with an open ended question (i.e. How can I help you?);
Ask the user to provide information about a specific attribute of the task (e.g. origin, destination, airline, departure time, etc.);
Retrieve data from the database according to the current user request (this action does not involve interaction with the user);
Output data to the user;
Ask the user to **release a constrain**;
Close the dialogue.

In order to simplify the learning task we chose a simplified state description. The state included three fields: the USER field represents the user request (only the keywords), the DATA field includes just the number of data tuples retrieved from the database, and the SYSTEM field records whether the actions **retrieve** or **output** have been already used.

We decided to use the following cumulated cost:

$$C = W_1 \cdot N + W_2 \cdot R + W_3 P + W_4 U$$

where N is the number of interactions, R is the cost of data retrieval (linear with the number of tuples retrieved), P is cost of data presentation (zero for a acceptable number of tuples and rapidly increasing with the number of tuples thereafter), and U is the user dissatisfaction cost that penalizes, with a fixed value, dialogues that fail to provide any flight information to the user.

The system starts with no knowledge of the task, i.e. actions are selected at random by the system at any step. A typical dialogue with an untrained system will look like the following:

S	RELEASE AIRLINE	*Do you want to choose another airline?*
U		What?
S	RETRIEVE	(6,893 flights retrieved)
S	CONSTRAIN DEPARTURE TIME	*When do you want to leave?*
U		Uh? I want to go to Boston
S	OUTPUT DATA	*There are 6,893 flights ....*
U		I don't understand, can you please show me the flights from San Francisco to Boston?
S	CLOSE DIALOGUE	*Thank you for using AT&T.*

Of course the total cost for this dialogue is quite high, especially due to the high cost of retrieval (all the database was retrieved here) and data output (6893 flights).

For learning the optimal strategy we used a *Monte Carlo* style reinforcement learning algorithm with *exploring starts*, as explained in [20].

With reinforcement learning it is not possible to train a system using a fixed corpus of interaction, since the system is learning while interacting, and the strategy is changing from iteration to iteration. Also it is not realistic to use real users for the learning stage that can be very long (thousands of interactions) and annoying (see the previous dialogue). So we decided to use a stochastic user model [17] designed to simulate the interaction. The user model includes probability distributions over user's dimensions such as compliance in answering specific questions and number and type of unsolicited constraints provided in open-ended questions. Those parameters were estimated on the ATIS corpus.

With the user model interacting with the reinforcement learning algorithm for thousands of interactions the system was able to learn the following optimal strategy:

- start the dialogue by **greeting** (the probability of getting more information from the user is higher with greeting rather than a specific question),
- ask **constraining** questions until the origin, destination, and airline are specified,
- **retrieve** data from the database.
- if the resulting data set is empty (because the query was over-constrained), **release** the airline constraint and **retrieve** again
- if there are too many flights in the data set, ask to **constrain** the departure time and retrieve again.
- if at any point during the dialogue the retrieved data set has a reasonable number of flights, the data is presented to the **output** and the dialogue is **closed**.

The following is an example of dialogue with the learned strategy:

S	GREETING	*This is the AT&T flight information system, how may I help you?*
U		I want to go to Boston
S	CONSTRAIN ORIGIN	*What is your departure airport?*
U		San Francisco.
S	CONSTRAIN AIRLINE	*Which airline?*
U		Delta.
S	RETRIEVE	(30 flights retrieved)
S	CONSTRAIN DEPARTURE TIME	*When do you want to leave?*
U		Late in the afternoon.
S	RETRIEVE	(3 flights retrieved)
S	OUTPUT DATA	*Flight ... leaves at ..., flight ... leaves ...*
U		Thanks.
S	CLOSE DIALOGUE	*Thank you for using AT&T.*

This strategy is the same as the one that was designed by hand and it is shown in Figure 5. This experiments demonstrates that it is possible build systems that are able to learn an optimal strategy based only on the enumeration of the possible actions and on the definition of reasonable costs.

Summary

We discuss in this paper the evolution of spoken language dialogue systems starting from the user-initiated model, where the user asks questions and the system provides only the most reasonable answer in the context of the previous interaction. The following step in the evolution of dialogue system is mixed initiative. In this model, the system can take the initiative at any time during the conversation in order, for instance, to gather additional information from the user for a more effective achievement of the goal. We described a general architecture for implementing mixed initiative systems, based on the definitions of system state, dialogue actions, and strategy. This architecture naturally leads to a formal definition of a

dialogue system in terms of a Markov decision process, and to the possibility of learning the dialogue strategy while interacting. We show that learning is possible even in complex applications, and that reinforcement learning algorithms can find reasonable strategies that are optimal for given costs.

REFERENCES

[1] Pieraccini, R., Levin, E., "Stochastic Representation of Semantic Structure for Speech Understanding," Speech *Communication*, Vol.11 pp. 283-288, 1992.

[2] MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus," *Proc. Of Fifth Darpa Workshop on Speech and Natural Language*, Harriman, NY, Feb 1992.

[3] Pieraccini, R., Levin, E., "A learning approach to natural language understanding," in *Speech Recognition and Coding, New Advances and Trends*, NATO ASI Series, Springer, 1993.

[4] Pereira, F., Riley, M. D., Sproat, R., "Weighted rational transductions and their application to human language processing," ARPA Human Language Technology Workshop, Princeton, NJ, March 1994.

[5] Levin, E., Pieraccini, R. "CHRONUS, The Next Generation," Proc. of *1995 ARPA Spoken Language Systems Technology Workshop*, Austin Texas, Jan. 1995.

[6] Bocchieri, E.L., Riccardi, G. Anantharaman, J., "The 1994 AT&T ATIS CHRONUS Recognizer," Proc. of *1995 ARPA Spoken Language Systems Technology Workshop*, Austin Texas, Jan. 1995.

[7] Riccardi, G., Bocchieri, E., Pieraccini, R., "Non Deterministic Stochastic Language Models for Speech Recognition," Proc. of ICASSP 95.

[8] Kaelbling, L. P., Littman, M. L., Moore, A. W., "Reinforcement Learning: A Survey," in *Journal of Artificial Intelligence Research*, No. 4, pp. 237-285, May 1996.

[9] Sadek, M., D., et. al. "Effective Human-Computer Cooperative Spoken Dialogue: the AGS Demonstrator," in *Proc. of ICSLP 96*, Philadelphia, PA, Oct. 96.

[10] Gorin, A. L., Parker, B. A., Sachs, R. M. and Wilpon, J. G., ``How may I help you?," Proc. Interactive Voice Technology for Telecommunications Applications (IVTTA), Oct. 1996, pp. 57-60.

[11] Kamm, C., Narayanan, S., Ritenour, R. and Dutton, D., "Evaluating spoken dialog systems for telecommunication services," *Proc. Eurospeech97*, Greece, Sept. 1997, pp. 2203-2206.

[12] Walker, M., A., Hindle, D., Fromer, J., Di Fabbrizio, G., Mestel, C., "Evaluating Competing Agent Strategies fro a Voice Email Agent," *Proc. Eurospeech97*, Greece, Sept. 1997, pp. 2219-2222.

[13] Pieraccini, R., Levin, E. and Eckert, W., "AMICA: the AT&T Mixed Initiative Conversational Architecture," *Proc. Eurospeech 97*, Rhodes, Greece, Sept. 1997, pp 1875-1878

[14] Levin, E. and Pieraccini, R., "A stochastic model of computer-human interaction for learning dialogue strategies," Proc. of EUROSPEECH 97,Rhodes, Greece, Sept. 1997.

[15] Zue, V., "Conversational Interfaces: Advances and Challenges," Proc. of EUROSPEECH 97,Rhodes, Greece, Sept. 1997.

[16] Walker, M. A., Littman, D. J., Kamm, C. A., Abella, A., "PARADISE: A Framework for Evaluation of Spoken Dialogue Agents," in Proc. *35-th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, 1997.

[17] Eckert, W., Levin, E. and Pieraccini, R., "User modeling for spoken dialogue system," ASRU '97, Proc. of *1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, Dec.1997.

[18] Levin, E., Pieraccini, R., Eckert, W., "Learning Dialogue Strategies within the Markov Decision Process Framework," ASRU '97, Proc. of *1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, Dec.1997.

[19] Gorin, A. L., Riccardi, G. and Wright, J. H., ``How may I help you?," Speech Communication, To appear.

[20] R. S. Sutton, A. G. Barto, "Reinforcement Learning, an Introduction," MIT Press, 1998.

Reconnaissance II

Reconnaissance des Phonèmes par Réseaux de Neurones et Normalisation Temporelle: Application aux Consonnes Pharyngales et Glottale Arabes.

A. Amrouche[1], M. Debyeche[1], A. Adoul[1], K. Amrouch[1], J.M. Rouvaen[2]

[1]Lab. Signaux et Systèmes, Institut d'Electronique,
USTHB, B.P. 32, Bab Ezzouar, 16 111, Alger, Algeria.
tel: (213)2515575, fax:(213)02515039, e-mail: usthbele1@ist.cerist.dz
[2]Dépt. O.A.E, U.M.R C.N.R.S 9929,
Institut d'Electronique et de Microélectronique du Nord, UVHC
Le Mont Houy, B.P. 311, 59304, Valenciennes cedex, France.
tel: 33 327141238, fax: 33 327141189, e-mail: rouvaen@ univ-valenciennes.fr

Résumé

Dans cet article, nous proposons une application des modèles connexionnistes pour l'identification des consonnes pharyngales et glottale arabes, en vue de la reconnaissance de parole. Un réseau de neurones (ANN), basé sur l'algorithme de rétro-propagation du gradient (B.P) assure la classification des phonèmes. Le vecteur d'entrée de ce réseau est constitué par les coefficients cepstraux (Mfcc) extraits au cours de la phase de prétraitement acoustique des phonèmes. L'originalité de notre travail réside dans une normalisation temporelle des phonèmes, basée sur des algorithmes d'interpolation, et une architecture de réseau totalement paramétrée. Les résultats montrent une amélioration du taux de reconnaissance pour des consonnes qui présentent des difficultés de segmentation telles que les pharyngales.

Mots clé: Identification des phonèmes, Retropropagation du gradient (BP), Normalisation temporelle.

Abstract

In this paper, an artificial neural network (ANN) has been used as a phonetic classifier for arabic phoneme recognition. Back-Propagation algorithm (BP) is used to train and test. The phoneme durations are firstly normalized. Input vectors for the neural network are the Mel Frequency Cepstrum Coefficients which are generated during the acoustic preprocessing phase. The originality of our work is a temporal normalisation based on interpolation algorithms. The recognition results of arabic back consonants indicate that the normalisation used may be a solution to phonetic identification based on ANN.

Keywords: Speech Recognition, Phoneme Identification, Back-Propagation, Temporal Normalisation.

1. Introduction

Dans l'approche analytique, l'identification des phonèmes constitue la partie essentielle des systèmes de reconnaissance de parole. Ce qui suppose au préalable une séparation des phonèmes en macro-classes phonétiques telles que les voyelles, les plosives, les fricatives, les nasales,... L'identification des phonèmes au sein d'une même classe est un problème difficile à résoudre. Les méthodes basées sur l'intelligence artificielle (I.A) nécessite un nombre de règles très élevé, dont la mise en oeuvre s'avère souvent laborieuse, tout en faisant appel à une expertise phonétique.

L'approche connexionniste [Wid90] consiste à traiter ce problème par des mémoires associatives. Ainsi, à chaque phonème correspond un ensemble de représentations mémorisées lors de la phase d'apprentissage sous forme de matrice de poids synaptiques. Lors de la phase de reconnaissance, le phonème inconnu est présenté à l'entrée du réseau qui tente alors un rapprochement avec les représentations faites lors de la phase d'apprentissage.

Le problème consiste donc à définir le type et l'architecture du réseau, en particulier ses dimensions, mais surtout ses vecteurs d'entrée, car ceux-ci sont directement liés aux phonèmes. Un intérêt particulier doit aussi être accordé à la base de données acoustique constituée des phonèmes préalablement segmentés et étiquetés.

2. Le modèle proposé

Le modèle proposé comporte trois parties complémentaires: la normalisation temporelle des phonèmes, le prétraitement acoustique et enfin, l'utilisation d'un réseau de neurones, dont le vecteur d'entrée est constitué des coefficients cepstraux et éventuellement d'indices acoustiques, pour identifier les phonèmes.

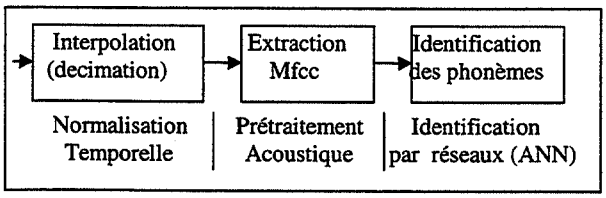


Fig. 1. Architecture du modèle d'identification des phonèmes.

2.1. Normalisation Temporelle des Phonèmes

La durée des phonèmes, qui est l'un des paramètres prosodiques, varie en fonction du contexte segmental et suprasegmental [Amr96]. Cette instabilité va donner lieu, pour un même phonème, à un nombre de coefficients cepstraux qui ne sera pas constant. Ce qui va nécessairement poser le problème de la taille du réseau, ainsi que de son mode de fonctionnement. Dans la littérature, plusieurs alternatives ont été proposées [Latr94, Brun94]. Nous proposons dans cette article une méthode originale qui consiste en une normalisation, en aval, des entrées. La normalisation temporelle a pour but d'ajuster tous les phonèmes à un vecteur de taille fixe, pour avoir un nombre de coefficients Mfcc constant.

2.2. Prétraitement Acoustique

La phase de prétraitement acoustique a consisté en l'extraction de coefficients cepstraux des phonèmes préalablement segmentés et normalisés. Pour se rapprocher du modèle de l'oreille, ces coefficients cepstraux ont été calculés dans une échelle Mel à l'aide d'un ensemble de filtres. Dans notre cas, les coefficients Mfcc ont été obtenus par une transformée en cosinus sur le logarithme de la distribution spectrale d'énergie.

2.3. Le Réseau de Neurones

L'architecture du réseau d'identification des phonèmes est donnée par la figure 2. Le problème de l'instabilité temporelle des phonèmes étant résolu, nous avons opté pour un réseau fixe, dont les dimensions sont totalement paramétrées. Le réseau est basé sur la rétropropragation du gradient de l'erreur [Wid90].

En phase d'apprentissage, on fournit à la fois l'entrée et la sortie connue. Nous disposons de P paires de vecteurs (x_p, yd_p), composant la base d'apprentissage (B.A).
Soit le vecteur d'entrée $x_p = (x_{p1}, ..., x_{pN})$ appliqué à l'entrée du réseau, N étant la dimension de ce vecteur, dont les éléments sont les coefficients Mfcc et éventuellement d'autres indices acoustiques.

La sortie de la $j^{ième}$ cellule de la couche cachée est:

$$s_{pj}^{(2)} = f\left(\sum_{i=1}^{N} w_{ji}^{(2)}.x_{pi} + q_j^{(2)}\right) \quad (1)$$

w_{ji}: matrice des poids de cette cellule avec la $i^{ième}$ entrée (provenant de la couche d'entrée).
q: seuil ou biais.
f(): fonction sigmoïde

$$f(x) = \frac{1}{1+e^{-x}} \quad (2)$$

Cette sortie sera l'entrée de la $k^{ième}$ cellule de la couche de sortie finale.

La mise à jour des poids synaptiques s'obtient par la relation:

$$w_{kj}(t+1) = w_{kj}(t) + \mu.\delta_{pk}.s_{pj} \quad (3)$$

μ: paramètre d'apprentissage.

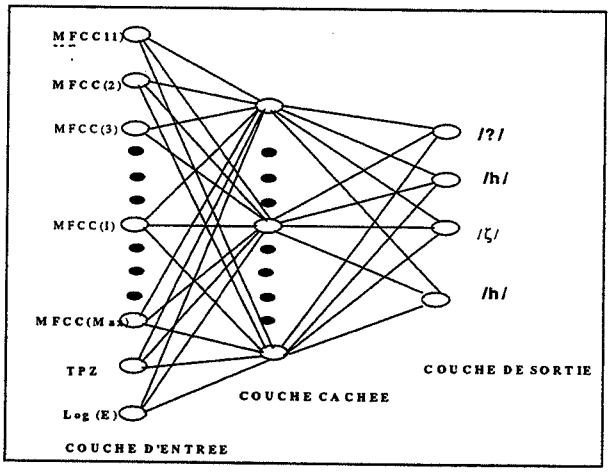


Figure 2. Le réseau de neurones artificiels pour la reconnaissance des consonnes pharyngales.

2.4 La Base de Données Acoustiques

La base de données acoustiques est constituée par un ensemble de phonèmes /h/,/ħ/,/ʕ/ et /?/ segmentés à partir d'un corpus de 22 phrases répétées quatre fois par six locuteurs (4 masculins et 2 féminins). Elle est divisée en une base d'apprentissage B.A (3 répétitions parmi les 4) à utiliser au cours de la phase entraînement du système et en une base de test B.T, à utiliser pour l'évaluation. Cette dernière, constituée de la dernière répétition, a été complétée par des phonèmes segmentés à
partir de phrases lues par des locuteurs étrangers à la base d'apprentissage.

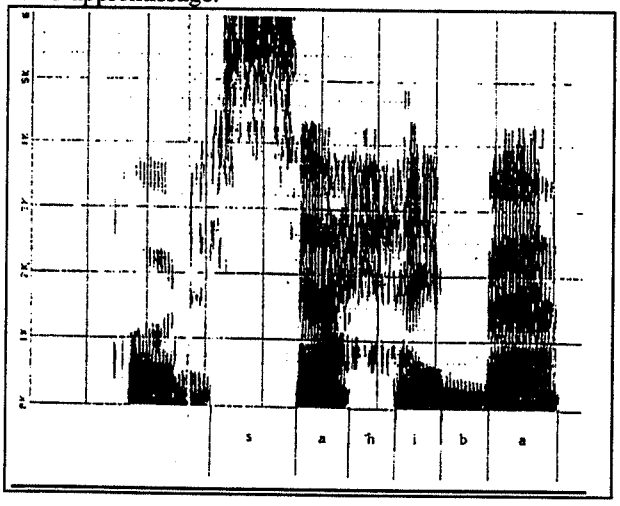


Figure 3: Segmentation de la consonne /ħ/ partir du mot saħiba.

3. Implémentation

3.1. Normalisation de la Taille des Phonèmes

L'algorithme de normalisation temporelle de la durée des phonèmes consiste à ajuster la durée des phonèmes à une durée type, caractéristique de la consonne considérée. Pour cela, un calcul statistique est fait sur toute la base d'apprentissage afin de déterminer la valeur moyenne de la durée du phonème, à travers le nombre d'échantillons qu'il comporte. Ensuite, tous les phonèmes de ce type sont ajustés à ce nombre soit par suppression, soit par duplication d'un nombre d'échantillons judicieusement calculés et situés, afin de ne pas altérer de manière significative le phonème original. Les échantillons à supprimer (ou à dupliquer) seront localisés de manière uniformément répartie sur l'ensemble du vecteur phonème.

3.1.1 Calcul de la durée moyenne D0 (Taille T0)
Soit une base d'apprentissage (B.A) comportant N phonèmes Pi à normaliser, 1<i< N.

Pi={s(j)}, j: rang de l'échantillon s(j), 1<j<Ti
Ti étant la dimension du vecteur phonème Pi, la taille moyenne T0 est calculée par :

$$T0 = \frac{\sum_{i=1}^{N} Ti}{N} \qquad (4)$$

$$D0 = T0 / Fech \qquad (5)$$

3.1.2 Normalisation des phonèmes à D0 (T0)
/*Soit Px un phonème à normaliser, Tx étant sa taille */

Si Tx ≤ T0 *alors faire* rajout de D=T0-Tx échantillons
Sinon faire suppression de S=Tx-T0 échantillons

procédure rajout
m= Tx modulo (D)
Tant que Tx<T0 *faire*
n=1;
dupliquer l'échantillon s(m.n);
Tx=Tx+1;
n=n+1;
fin
récupérer fichier phonème normalisé de dimension T0;
fin;

La procédure suppréssion d'échantillons obeit au même principe de répartition uniforme des échantillons à traiter.

3.1.3Aspect Pratique de la Normalisation Temporelle
Pour des raisons de commodité liées aux phases extraction des coefficients Mfcc et classification par réseau de neurones, la dimension T0 a été prise comme un multiple d'une trame de 256 échantillons.

3.2. Extraction des Coefficients Cepstraux

Le calcul des coefficients Mfcc (12 par trame), a nécessité l'implémentation des modules suivants :

-Fenêtrage et préaccentuation avec $H(z)=1-0.96z^{-1}$.
-Calcul FFT sur 256 points avec recouvrement de moitié
-Calcul de l'énergie
-Initialisation de 24 filtres échelonnés suivant une loi logarithmique(Mel).
-Filtrage, Calcul et Stockage des coefficients Mfcc.

3.3. Réseau de Neurones

Le réseau de neurones simulé est basé sur de retropropagation de gradient [Wid90], qui s'adapte bien au traitement des signaux [Fre92, Her94]. Bien que le nombre de couches cachées, ainsi que le nombre de cellules internes à ces couches, sont implémentés sous forme des de paramètres, nous n'avons retenue que la configuration à une couche cachée. En effet, les expériences que nous avons menées avec un nombre de couches cachées multiple ont montré une altération des taux de reconnaissance.

Les entrées du réseau sont constituées des coefficients Mfcc. Pour notre application, la couche de sortie comporte quatre cas possibles correspondant aux phonèmes: /h/,/ħ/,/ɕ/, et /?/. Pour faire ressortir les confusions, ces consonnes ont été codées sur quatre bits.

Notons toutefois que les nombres d'entrées et de sorties sont implémentés sous forme de paramètres afin de faciliter d'une part, la prise en compte d'autres indices acoustiques tels que le taux de passage par zéro (TPZ) et l'énérgie (log E), d'autre part permettre une adaptation du réseau à un nombre de phonèmes plus élevé

4. Expériences de reconnaissance

4.1 Expérience 1

nombre de cellules d'entrées: 84
nombre de cellules de sorties: 4
nombre de couches cachées: 1 avec 15 cellules
paramètre d'apprentissage: μ=0.25
seuil de l'erreur: $\varepsilon = 10^{-3}$
mémentum: α=0.1

Tableau 1. Taux de reconnaissance avec B.A.

phonème	total	reconnus	non reconnus	taux (%)
/?/	96	81	15	84,37
/ħ/	75	69	6	92
/ɕ/	90	74	16	82,22
/h/	61	60	1	98,36

Tableau 2. Taux de reconnaissance avec B.T

phonème	total	reconnus	non reconnus	taux (%)
/?/	51	33	18	64,70
/ħ/	71	53	18	74,64
/ɕ/	77	50	27	64,93
/h/	69	30	9	76,92

4.2 Expérience 2

nombre de cellules d'entrées: 84
nombre de cellules de sorties: 4
nombre de couches cachées: 1 avec 30 cellules
paramètre d'apprentissage: μ=0.1
seuil de l'erreur: $\varepsilon = 10^{-3}$
mémentum: α=0.8

Tableau 3. Taux de reconnaissance avec B.A.

phonème	total	reconnus	non reconnus	taux °/°
/ʔ/	96	94	2	97,91
/ħ/	75	70	5	93,33
/ç/	90	84	6	93,33
/h/	61	56	5	91,80

Tableau 4. Taux de reconnaissance avec B.T.

phonème	total	reconnus	non reconnus	taux °/°
/ʔ/	26	20	6	76.92
/ħ/	36	26	12	68,42
/ç/	42	19	23	45,23
/h/	19	13	6	68,42

5. Résultats et Discussions

Le taux de reconnaissance des phonèmes faisant partie de la base B.A. est nettement supérieur au taux de reconnaissance des consonnes faisant partie de la base B.T. Nous remarquons également que la qualité de la segmentation à une incidence majeure sur le taux de reconnaissance. Ainsi, pour la consonne /ç/, très difficile à segmenter, le taux de reconnaissance avec la base de test chute de manière significative. Pour cette consonne, le taux plus élevé obtenu avec la base d'apprentissage est dû au fait de l'apprentissage du réseau.

Une comparaison des résultats précédents avec ceux du tableau 5 montre que le taux de reconnaissance des phonèmes est amélioré par rapport à un système basé sur l'intelligence artificielle [Deb96].

Tableau 5: Taux de reconnaissance avec (I.A).

phonème	total	reconnus	taux (%)
/ʔ/	155	107	70
/ħ/	97	80	83
/ç/	98	64	65
/h/	48	31	65

Le problème majeur qui se posait pour une généralisation à une base de données plus étendue est le problème de la longueur du segment phonémique, donc du nombre de trames, qui varie avec la vitesse d'élocution des locuteurs. Nous pensons avoir partiellement résolu ce problème au sein d'une même classe de phonèmes avec l'introduction de la normalisation de la longueur des segments phonétiques.

Conclusion

La normalisation en aval de la longueur des segments permet l'utilisation d'un réseau de dimensions stables plus facile à mettre en oeuvre. Les résultats obtenus avec l'application à des consonnes particulières de l'arabe, parmi les plus difficiles à segmenter, montrent que cette méthode est bien adaptée à l'identification des phonèmes.
Pour une utilisation en reconnaissance de parole, cette architecture suppose, cependant, une segmentation et une classification préalables des phonèmes en classes homogènes. Insérée dans un tel ensemble, nous estimons que notre méthode peut palier aux incertitudes liées aux difficultés de segmentation
Une adaptation de cette méthode à la reconnaissance de mots isolés, tels que les chiffres, est en cours.

Références

[Wid90]
Widrow B, Lehr M.A [1990], "30 Years of Adaptative Neural Networks: Perceptron, Madaline, and Backpropagation", Proc. of I.E.E.E., Vol. 78, N° 9, 1990, pp.1415-1442.

[Fre92]
Freeman J.A, Spakura D.A [1992], *Neural Networks: Algorithms, Applications and Programming Technics*, Addison-Wesley Publishing Company, 1992.

[Her94]
Herault H, Jutten C [1994], *Réseaux de Neurones et Traitement de Signal*, Edition Hermes, Paris, 1994.

[Amr96]
Amrouche A, Boudraa B, Rouvaen J.M[1996], "Modèle de prédiction des durées des phonèmes pour la synthèse automatique de la parole". Actes 2ème COMAEI'96, Alg. Jour. of Technology, Vol. N°2, dec. 1996, pp 12-17.

[Lat94]
Latrucci L, Bellesi G, Gori M, Soda G [1994], "Architecture for Speaker Indepedent Phoneme Recognition", Neural Network for Signal Processing IV, Proc. of the 1994 IEEE Workshop, pp.309-318.

[Bru94]
Brunet P.T, Pandya A.S, Pinera C.V [1994]," Artificial Neural Network for Phoneme Recognition", Neural Network for Signal Processing IV, Proc. of the 1994 IEEE Workshop, pp.4473-4477.

[Deb96]
Debyeche M, Hatton J.P [1996],"Knowledge Based Approch for Arabic back Consonant Recognition in Continuous Speech", First KFUPM Workshop on Information and Computer Science, Dhahran, Saudi Arabia, 1996.

Modèles hybrides et reconnaissance de la parole continue indépendante du locuteur en français

O. Deroo, C. Ris, F. Malfrere[1] & T. Dutoit

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: deroo,ris,malfrere,dutoit@tcts.fpms.ac.be

ABSTRACT

Nous rapportons dans ce papier les résultats obtenus avec un modèle hybride HMM/MLP qui combine les réseaux de neurones et les modèles de Markov cachés pour la reconnaissance de la parole continue en français sur la base de données BREF-80. Ce corpus n'étant pas manuellement segmenté, nous décrivons une nouvelle méthode basée sur l'alignement du signal de parole avec un signal synthétique de bonne qualité pour générer une première segmentation afin d'initialiser le processus d'entraînement des modèles. Nous avons obtenus avec notre système de base un taux de reconnaissance de 75% au niveau du phonème (test réalisé sur 144 phrases énoncées par 8 locuteurs n'ayant pas participé à l'entraînement), ce qui est l'un des meilleurs taux rapporté dans la littérature sur ce corpus particulier [5]. Des expériences préliminaires au niveau du mot utilisant des vocabulaires de tailles 1K, 3K, 13K et 65K sont aussi rapportées. Toutes les expériences décrites dans ce papier ont été réalisées avec le logiciel STRUT [10] [1] et le décodeur grand vocabulaire NOWAY [9].

1. INTRODUCTION

Des avancées significatives dans le domaine de la reconnaissance de la parole continue grand vocabulaire indépendante du locuteur ont été réalisées ces dernières années. Les modèles de Markov cachés (HMM) sont de nos jours, l'approche la plus utilisée en reconnaissance de la parole. Pourtant, ils souffrent de nombreuses hypothèses contraignantes parmi lesquelles le fait que les vecteurs acoustiques sont supposés non corrélés ou encore l'hypothèse sur la distribution des densités de probabilités de chaque état HMM (distributions discrètes ou multigaussiennes). Ces hypothèses peuvent être contournées en utilisant un réseau de neurone couplé aux HMMs [6] [2]. Ce réseau de neurone (dans notre cas un MLP) estime les probabilités a posteriori utilisées par chaque état du HMM. Les modèles hybrides ont déjà été utilisés avec succès pour l'anglais britannique et américain. Dans ce papier, ce système est testé sur une base de données de parole continue en français :BREF. Quelques bons résultats sur cette base de données ont déjà été publiés [11] [5] utilisant des HMMs à densité de probabilité continue et des modèles de phonèmes dépendants du contexte. Nous comparerons ces résultats avec notre système de base, un système hybride indépendant du contexte.

2. LE CORPUS BREF-80

BREF-80 est un corpus de parole lue par 80 locuteurs. Les textes utilisés sont extraits du journal "Le Monde" afin de fournir un large corpus de parole continue grand vocabulaire (plus de 20.000 mots), et pour l'acquisition de connaissances acoustico-phonétiques sur le français. BREF-80 peut être efficacement utilisée pour entraîner différents types de modèles en français. Le dictionnaire de base représenté par 35 phonèmes a été obtenu à l'aide d'un phonétiseur par régles et a été manuellement vérifié. Le fichier de phonèmes a été étendu afin de tenir compte des liaisons potentielles du français. BREF-80 contient au total 5330 phrases énoncées par 80 locuteurs[3] qui ont été divisées en 3737 phrases pour l'entraînement (3363 pour l'entraînement du réseau et 374 pour la "cross validation" (utilisée pour adapter le taux d'apprentissage du MLP [6])), ce qui représente approximativement 9 heures de parole. Les données utilisées pour le test consistent en 144 phrases sélectionnées au hasard et énoncées par 8 locuteurs (4 hommes et 4 femmes)[4].

3. PARAMÈTRES ACOUSTIQUES

Les vecteurs acoustiques utilisés dans les expériences décrites dans ce rapport sont les PLP et log RASTA-PLP [7] ainsi que les cepstres avec soustraction cepstrale (Cepstral Mean Substraction CMS) [8] et les MFCC (Mel-scale Frequency Cepstral Coefficients). Ces paramètres sont calculés toutes les 10 ms sur une fenêtre d'analyse de 30 ms. L'ordre de l'analyse LPC a été fixé à 10. Ainsi le vecteur acoustique fourni à l'entrée de notre MLP est composé de 26 paramètres (les paramètres cepstraux, leurs dérivées premières ainsi que les dérivées première et seconde de l'énergie). De plus 9 trames de contexte acoustique sont utilisées à l'entrée du MLP, ce qui donne 234 noeuds d'entrées (9 trames de contexte correspondant à la configuration connue pour donner les meilleurs résultats). Enfin, une couche cachée de 1000 noeuds est utilisée. Le nombre de paramètres à entraîner est donc d'environ 270.000 (ce qui est largement inférieur à la plupart des

[1] Supporté par le F.R.I.A. (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture).

[2] Ces modèles sont couramment appelés modèles hybrides.

[3] Ce n'est qu'une petite partie du corpus total pour lequel 600 phrases ont été enregistrée par 120 locuteurs.

[4] Nous avons pris approximativement le même nombre de phrases que lors des expériences décrites dans [5].

systèmes décrits dans la littérature).

4. PROCÉDURE DE SEGMENTATION ET D'ENTRAÎNEMENT DES MODÈLES

Dans tous les cas, nous avons utilisé un entraînement Viterbi embedded. Ce type d'entraînement requiert une segmentation phonétique de la base de données. Comme aucune segmentation n'est fournie avec BREF, nous avons généré une segmentation initiale pour amorcer le processus d'entraînement par une nouvelle méthode [4] basée sur l'utilisation d'un synthétiseur vocal de grande qualité (MBROLA [3]). Ce synthétiseur est en fait utilisé pour générer (à partir de la transcription phonétique) un signal de parole synthétique qui est aligné sur le signal de parole original. Ainsi le processus d'alignement est réduit à un simple algorithme DTW ("Dynamic Time Warping"). Le processus global de segmentation peut être visualisé sur la figure 1.

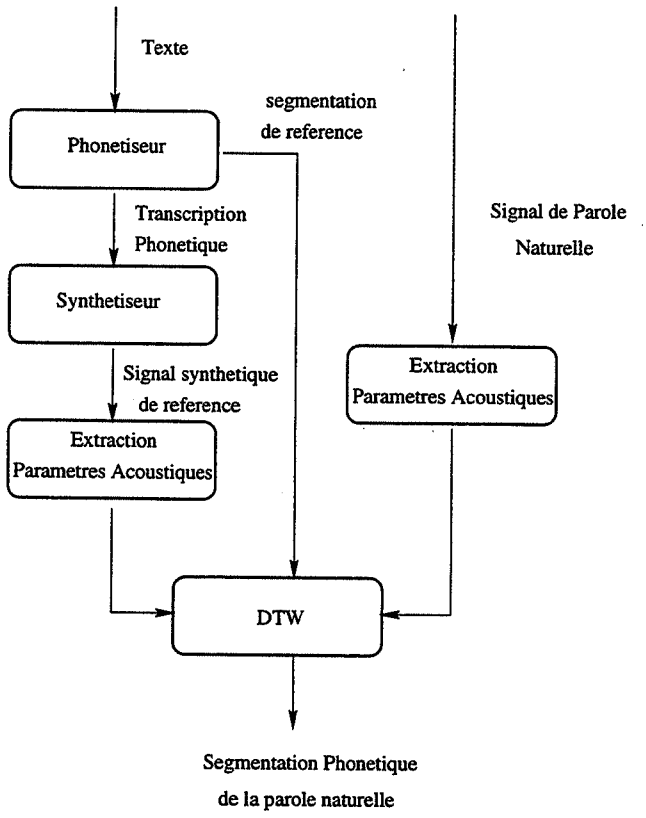


Figure 1: *Alignement Phonétique avec un TTS.*

En comparaison avec les méthodes classiques HMMs, cette méthode perd apparemment l'indépendance par rapport au locuteur puisqu'une seule voix est utilisée pour générer le signal acoustique. Mais la connaissance précise des limites entre les phonèmes suffit largement à contrebalancer cet effet. Notons qu'en pratique, nous avons observé que l'utilisation de deux voix distinctes (une pour les femmes et une pour les hommes) améliore sensiblement la pertinence de la segmentation. Fabrice Malfrère et Thierry Dutoit [4] ont reporté un taux d'erreur de segmentation d'environ 8% (si on considère qu'une déviation temporelle de plus 50 ms implique que le segment est faux). L'avantage principal de cette méthode réside dans le fait qu'aucune base d'entraînement n'est requise. Il suffit d'avoir la transcription phonétique. La segmentation obtenue par cet algorithme est indubitablement bien plus réaliste qu'une segmentation linéaire particulièrement lorsque les phrases sont assez longues (ce qui est le cas pour BREF. Il y a fort à parier que le processus d'entraînement basé sur une segmentation linéaire aurait eu de grandes difficultés à converger. Cette segmentation initiale (facilement et rapidement obtenue) a été utilisée pour initialiser le processus d'entraînement et a été affinée par itérations successives. Comme le prouve les taux de reconnaissance au niveau de la trame dans le tableau 1, la procédure suivie pour cette base de données nous a permis d'entraîner rapidement un bon système en français.

	Trames	RASTA	PLP	CMS	MFCC
Train	2400000	79.6%	82.0%	82.4%	82.4%
Cross	270000	77.0%	80.2%	79.6%	79.6%

Table 1: *Taux de reconnaissance au niveau de la trame acoustique avec un système hybride HMM/ANN entraîné sur des paramètres log-RASTA-PLP, PLP, CMS et MFCC.*

Ainsi l'utilisation d'un synthétiseur vocal de grande qualité est un moyen efficace de générer une première segmentation d'une base de données qui peut être utilisée comme segmentation initiale dans le processus d'entraînement des systèmes de reconnaissance.

5. RECONNAISSANCE DE PHONÈMES SUR BREF-80

Nous avons réalisé une expérience de reconnaissance de phonèmes sur cette base de données particulière afin de tester notre modèle hybride. Le système de base utilise des phonèmes indépendants du contexte (35) décrits dans le tableau 2 (SAMPA [12]). Chaque modèle HMM est composé de N états identiquement distribués où N est la durée moyenne de chaque phonème divisé par deux (on impose ainsi une durée minimale de N trames acoustiques lors de son passage dans le modèle) (Voir figure 2). Il est à noter aussi qu'aucune grammaire n'a été utilisée dans les expériences décrites dans ce paragraphe.

Figure 2: *Un modèle de phonème indépendant du contexte avec N états partageant les mêmes paramètres statistiques.*

Comme vous pouvez le constater sur le tableau 3, les taux de reconnaissance obtenus sont assez impressionnant. Un taux de reconnaissance de 75.0% a été observé ce qui correspond au meilleur taux jamais publié à notre connaissance sur cette base de données [5].

Vowels		Fricatives	
i	idée	f	feu
e	ému	s	soeur
E	maison	S	chanter
a	alarme	v	vent
O	obstacle	z	zéro
o	beau	Z	jardin
u	coupable	Plosives	
y	punir	p	patte
2	creuser	t	tarte
9	peur	k	coup
@	petite	b	robe
Nasal Vowels		d	dire
e~	lundi	g	gauche
o~	rondeur	Nasals	
a~	temps	m	madame
Semi Vowels		n	nous
j	piétiner	N	campagne
w	soir	Others	
H	huile	G	boeing
l	long	-	silence
R	rond		

Table 2: *Les 35 phonèmes utilisés.*

log-RASTA	PLP	CMS	MFCC
69.2%	72.9%	63.8%	75.0%

Table 3: *Taux de reconnaissance de phonèmes utilisant un modèle hybride et différents types de paramètres.*

6. RECONNAISSANCE DE LA PAROLE CONTINUE

Le système de base étant performant au niveau de la trame acoustique et du phonème, nous avons décidé de le tester au niveau du mot. Nous avons utilisé 4 tailles de dictionnaires différentes :

- Un dictionnaire de 1.000 mots contenant les mots uniquement contenus dans les phrases de tests.
- Un dictionnaire de 3.000 et 13.000 mots contenant les mots du test set complétés avec les mots les plus courants.
- Un dictionnaire de 65.000 mots.

Ces dictionnaires ont été obtenus par l'intermédiaire d'un phonétiseur par régles. Les résultats obtenus avec notre système hybride HMM/ANN de base sont rapportés dans les tableaux 4 et 5. Il est à noter que nous avons utilisé deux modèles de langage (bigramme et trigramme) estimés sur les textes du monde (années 90-91-92 soit environ 80 millions de mots) en utilisant les outils de modèlisation du CMU-Cambridge. Nous avons pris garde à ce que les textes des données de tests ne figurent pas dans les textes utilisés pour la construction des modèles de langage. Les perplexités respectives de ces modèles de langage sont de 151.6 pour le bigramme et 94.4 pour le trigramme.

Nous ne rapportons ici que les meilleurs résultats qui ont été obtenus au niveau du mot, c'est pourquoi uniquement les résultats avec les paramètres acoustiques PLP sont donnés.

Taille du dictionnaire	Ins.	Del.	Subs.	Erreur
1K	2.9%	5.0%	12.9%	20.8%
3K	3.4%	5.2%	14.0%	22.5%
13K	3.1%	6.2%	17.4%	26.7%
65K	2.7%	6.3%	18.9%	28.0%

Table 4: *Taux d'erreur au niveau du mot pour des vocabulaires de 1K, 3K, 13K, 65K avec un modèle hybride HMM/ANN et un modèle de langage bigramme.*

Taille du dictionnaire	Ins.	Del.	Subs.	Erreur
1K	2.3%	4.1%	9.5%	15.9%
3K	3.0%	4.6%	13.8%	19.8%
13K	3.1%	5.6%	15.3%	24.1%
65K	2.5%	6.0%	16.5%	25.0%

Table 5: *Taux d'erreur au niveau du mot pour des vocabulaires de 1K, 3K, 13K, 65K avec un modèle hybride HMM/ANN et un modèle de langage trigramme.*

Les résultats présentés ici sont les premiers obtenus avec un modèle hybride HMM/ANN sur cette base de données. Il est à noter que nous avons été confrontés à de nombreux problèmes dus au langage utilisé (le français). Les systèmes de reconnaissance souffrent beaucoup des liaisons entre les mots (ce qui arrive fréquemment en français) et du fait que certains phonèmes sont par moment prononcés par certains locuteurs et omis par d'autres. Ce problème a été partiellement résolu en utilisant pour chaque mot des prononciations multiples (modélisant au mieux les différentes prononciations dans le cas d'ommission de phonèmes ou de liaisons).

De plus en français, les homophones sont un réel problème. Par exemple les phrases "c'est à peine" et "sept à peine", phonétiquement identiques, sont toutes les deux sémantiquement valides mais occasionneraient des erreurs lors du calcul des taux de reconnaissance. Gauvain [5] reporte un taux d'homophone de 30% en français ce qui est largement supérieur à celui de l'anglais par exemple (3% pour TIMIT).

Les taux de reconnaissance au niveau du phonème ainsi que l'étude détaillée des erreurs commises lors du processus de reconnaissance montrent que la modèlisation acoustique des phrases est de très bonne qualité. Il nous reste cependant encore plusieurs points à explorer pour améliorer les taux de reconnaissance au niveau du mot.

Dans un premier temps, nous allons tenter de travailler sur les modèles de langage et sur les lexiques utilisés. Nous pourrions aussi tenter d'entrainer des modèles de phonèmes dépendants du contexte (pour lesquels nous avons déjà observé des améliorations significatives du taux de reconnaissance [2]). Néanmoins, le peu de données d'entraînement que nous possédons (environ 5500 phrases) risque d'être un facteur limitatif de l'amélioration que nous pourrions observer.

7. CONCLUSIONS ET PERSPECTIVES

Ces premières expériences ont permis de fixer les performances de base pour notre système hybride HMM/ANN sur BREF. Les taux de reconnaissance au niveau du phonème sont apparemment les meilleurs jamais rapportés dans la littérature avec un système assez simple (peu de paramètres, modèles de phonèmes indépendant du contexte, aucune grammaire). De plus, nous avons proposé une nouvelle méthode basée sur l'utilisation d'un synthétiseur vocal pour obtenir une meilleure segmentation phonétique initiale du signal de parole, permettant ainsi une accélération de la convergence de l'algorithme d'entraînement des modèles. Les taux de reconnaissance au niveau du mot sont moins bons que nous ne l'espérions. Nous prévoyons d'explorer les voies suivantes pour améliorer notre système :

- utilisation de modèles dépendants du contexte avec notre système hybride.
- Rescoring des N meilleures séquences de mots par un modèle de langage plus performant ou par des règles grammaticales.

En effet, l'étude détaillée des N meilleures séquences montre que la bonne phrase est souvent parmi les 100 meilleures. Nous espérons donc pouvoir utiliser d'autres informations pour rescorer ces N meilleures phrases et améliorer ainsi le taux de reconnaissance au niveau du mot. Ce système devrait bientôt être intégré dans la prochaine version de l'outil de reconnaissance de parole continue grand vocabulaire ABBOT [9].

BIBLIOGRAPHIE

[1] J.-M. Boite, F. Bataille, O. Deroo, S. Dupont, V. Fontaine, C. Ris, and L. Zanoni, "Strut : Un logiciel complet pour l'entraînenment et la reconnaissance de la parole," in *Proc. Premieres Journees Scientifiques et Techniques FRANCIL*, pp. 41–44, Apr. 1997.

[2] S. Dupont, C. Ris, O. Deroo, and V. Fontaine, "Context independent and context dependent hybrid hmm/ann systems for vocabulary independent tasks," *European Conference On speech communication and Technology*, pp. 1947–1951, September 1997. Patras, Greece.

[3] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van Der Vreken, "The mbrola project : Toward a set of high quality speech synthesizers free of use for non commercial purposes," *International Conference on Speech and Language Processing*, 1996. Philadelphia.

[4] F. Malfrere and T. Dutoit, "High-quality speech synthesis for phonetic speech segmentation," *European Conference On Speech Communication and Technology*, vol. 2, pp. 2631–2634, 1997.

[5] J. Gauvin and L. Lamel, "Speaker-independent phone recognition using bref," in *Proceedings of the DARPA Speech and Natural Language Workshop, Arden House, NY*, pp. 344–349, 1992.

[6] H. bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publisher, 1994.

[7] H. Hermansky. and N. Morgan, "Rasta processing of speech.," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.

[8] D. Van Hamme, G. Gallopyn, W. Van-Springel, B. D'Hoore, M. Butnaru and H. Bourlard, "Acoustic features comparison and robustness tests of a real-time recognizer on a hardware telephone line simulator," *Proceedings of International Conference on Speech and Language Processing*, 1994.

[9] M. Hohberg ,G.D. Cook ,S. Renals ,A.J. Robinson and R.S. Schechtman , "The 1994 abbot hybrid connectionist-hmm large vocabulary recognition system," *Spoken Language Systems Technology Wokshop*, pp. 170–176, January 1995.

[10] J.M. Boite ,O. Deroo ,V. Fontaine and C. Ris, "Step by step guide to using the speech training and recognition unified tool (STRUT)," *Internal Report*, http://tcts.fpms.ac.be/speech/strut/users-guide/users-guide.html.

[11] S.J. Young and al, "Multilingual large vocabulary speech recognition : the european SQALE project," *Computer Speech and Language*, pp. 73–89, 1997.

[12] University College London, *Multilingual Speech Input / Output Assesment Methodology and Standardisation*, 1988. Esprit Project 1541, Definition Phase Final Report.

Reconstruction de Données Manquantes pour la Reconnaissance Robuste de la Parole dans le Cadre des Systèmes Hybrides HMM/ANN

Stéphane Dupont[1]

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: dupont@tcts.fpms.ac.be

ABSTRACT

In this paper, we propose to use the missing features theory to allow the reconstruction of missing spectro-temporal features in the framework of hybrid HMM/ANN systems. A simple signal-to-noise ratio estimator allows to indicate the features that are unavailable or corrupted by noise (missing components). A limited number of multidimensional gaussian distributions is then used to reconstruct those missing components solely on the basis of the present features. The reconstructed vectors are then used as input to an artificial neural network estimating the HMM state probabilities. Continuous speech recognition experiments have been done on filtered speech. In this case, filtered components carry few or no information at all, and hence, should probably be ignored. The results presented in this paper illustrate this point of view. Experiments on noisy speech also confirm the interest of the proposed approach.

1. INTRODUCTION

Les recherches concernant la robustesse des systèmes de reconnaissance automatique de parole (robustesse aux bruits additifs, convolutifs...), se sont essentiellement focalisées sur des techniques de compensation au niveau des paramètres acoustiques (par exemple soustraction spectrale, filtrage RASTA...) [1, 8] ou sur des techniques de compensation des paramètres des modèles du signal (par exemple sur base d'un modèle de la perturbation [7]). Ces techniques ont pour but soit d'obtenir une version débruitée de la séquence de paramètres acoustiques, soit d'obtenir une version compensée (bruitée) des paramètres représentant les distribution de probabilités associées aux états des modèles de Markov cachés.

Cependant, il y a des cas où il est peut être préférable d'ignorer certains éléments des vecteurs de caractéristiques pour ne baser les décisions subséquentes que sur les éléments qui semblent les moins perturbés (on peut aussi envisager d'exploiter la redondance du signal pour obtenir une version reconstruite des éléments perturbés, sur base des éléments non perturbés). C'est le cas lorsque, suite à un filtrage du signal de parole, sa bande passante est différente de celle qui a été utilisée lors du développement du système. C'est également le cas lors de courtes interruptions du signal ou lorsque le niveau de bruit local (dans le plan spectro-temporel) est bien supérieur au niveau du signal utile. D'autre part, il apparaît que l'audition humaine est relativement insensible à des dégradations de ce type, ce qui tend à montrer qu'elle opère par distinction des portions spectro-temporelles fortement perturbées ou absentes, des autres portions. Des recherches récentes [6] ont tenté d'exploiter cette idée simple. Les résultats rapportés montrent qu'il est possible d'ignorer une part importante des éléments de la représentation spectro-temporelle (jusqu'à 90% dans certains cas) du signal sans pour autant détériorer significativement les performances du système de reconnaissance. On qualifie les éléments ignorés de **données manquantes**. Les éléments effectivement utilisés pour la classification sont appelés **données présentes**.

Dans cet article, nous envisageons la possibilité d'utiliser des techniques de données manquantes pour la reconstruction des vecteurs caractéristiques. Les vecteurs caractéristiques reconstruits sont alors utilisés par un système basé sur un réseau de neurones artificiel [3]. Comme illustré à la Figure 1, cette approche requiert également l'utilisation d'un module dont le but est d'indiquer les portions spectro-temporelles qui seront considérées comme manquantes par l'algorithme de reconstruction. Notons que le système multibande proposé dans [4] constitue une approche alternative à ce principe de reconnaissance partielle.

2. RECONSTRUCTION DE DONNÉES MANQUANTES

Les vecteurs d'observations x sont supposés indépendants et distribués de façon identique suivant une multi-gaussienne comportant K distributions gaussiennes multidimensionnelles. La i-ème distribution est déterminée par les paramètres suivants: w^i, le poids associé à la distribution, μ^i, sa moyenne et, C^i, sa matrice de covariance.

Certains des éléments de x étant considérés comme manquants, on peut réordonner x de la manière suivante:

$$x = (x_p x_m), \tag{1}$$

x_p correspond aux données présentes et x_m aux données manquantes. De la même façon, réordonnons les éléments des vecteurs de moyenne et des matrices de covariances

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture).

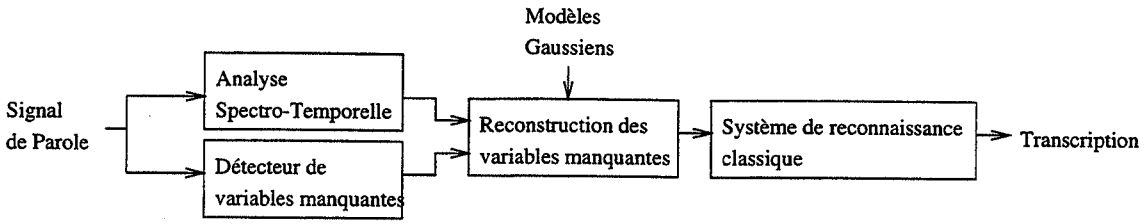


Figure 1: Système de reconnaissance automatique de parole utilisant la reconstruction de données manquantes.

caractérisant la distribution de x:

$$\mu^i = (\mu^i_p \mu^i_m), C^i = \begin{bmatrix} C^i_{pp} C^i_{pm} \\ C^i_{mp} C^i_{mm} \end{bmatrix} \quad (2)$$

Dans le cas qui nous occupe, nous souhaiterions reconstruire un vecteur complet à partir des données présentes. La reconstruction est effectuée en considérant la moyenne des données manquantes étant donné les valeurs des données présentes. Cette moyenne a une distribution gaussienne. La valeur reconstruite sera la moyenne de cette distribution, soit, pour la gaussienne i:

$$x^i_{m|p} = \mu^i_m + (C_{pm})^t (C_{pp})^{-1} (x_p - \mu_p) \quad (3)$$

Considérant la distribution multi-gaussienne, la valeur reconstruite est calculée selon la formule suivante:

$$x_{m|p} = \sum_{i=1}^{K} w^i \phi(x, \mu^i_p, C^i_{pp}) x^i_{m|p} \quad (4)$$

où w^i est le poids associé à la composante i de la distribution multi-gaussienne et $\phi(x, \mu^i_p, C^i_{pp})$ est la distribution gaussienne multidimensionnelle associée. Ce terme permet de pondérer les contributions des différentes gaussiennes en fonction de la position du vecteur de données présentes dans l'espace des paramètres.

Une alternative consisterait à ignorer les données manquantes pour ne calculer les probabilités associées aux différents états des modèles de Markov cachés que sur base de distributions marginales. Dans le cas où le classificateur est du type paramétrique (dans le cas de distributions multi-gaussiennes par exemple), c'est possible, bien que cela requiert une charge de calcul très importante. Il convient en effet d'inverser les matrices de covariance C_{pp} chaque fois que la décomposition du vecteur caractéristique en données manquantes et données présentes change. Dans le cas où le classificateur est du type réseau de neurone artificiel, c'est également possible mais très lourd. Il convient en effet d'avoir à sa disposition un réseaux de neurone pour chaque configuration de données présentes.

Les expériences décrites dans [6] concernent l'utilisation de distributions multi-gaussiennes. Les résultats présentés vont en faveur de l'utilisation de distributions marginales. L'utilisation d'une approche de reconstruction donne des résultats sensiblement moins bons. Cette approche comporte cependant plusieurs avantages. D'une part, on peut n'utiliser qu'un nombre restreint de gaussiennes pour la reconstruction, ce qui permet de limiter le nombre d'inversions de matrices, sans préjudice significatif pour le système de reconnaissance. D'autre part, elle permet d'obtenir des vecteurs reconstruits qui peuvent être utilisés par un système de reconnaissance classique. Il s'agira dans notre cas d'un système hybride HMM/ANN utilisant un perceptron multicouche comme estimateur des probabilités des classes phonétiques.

3. Expériences avec la base de données Numbers'93

Nous avons développé des systèmes de reconnaissance de nombres connectés en anglais, destinés à des applications téléphoniques à partir de la base de données Numbers'93 [5]. Nous avons utilisé 1534 phrases pour l'entraînement de nos systèmes (1400 pour ajuster les poids des réseaux de neurones artificiels et 134 dans un but de validation croisée) et 384 phrases pour tester les systèmes développés.

Le vocabulaire du corpus comprend 36 mots permettant de prononcer n'importe quel nombre inférieur à un million. Les transcriptions phonétiques de ces mots sont celles fournies par le dictionnaire phonétique *CMU 0.4*. Celui-ci utilise un sous-ensemble des phonèmes de la base de donnée TIMIT, en l'occurence les 32 phonèmes permettant de transcrire tous les nombres en anglais.

Nous avons développé un système de base utilisant un perceptron multicouche comme estimateur de probabilités. Nous avons utilisé comme paramètres représentatifs les énergies de bandes critiques filtrées par un filtre log-RASTA [8] (ainsi que les dérivées premières de ces paramètres), conférant d'excellentes performances de robustesse face aux caractéristiques du canal de transmission. Les entrées du réseau de neurones artificiel comportaient 9 vecteurs acoustiques consécutifs, fournissant une information contextuelle importante au système. Le taux d'erreur au niveau du mot était de **13.6%**. Remarquons que ces performances sont significativement inférieures à celles obtenues en utilisant des paramètres log-RASTA-PLP, c'est-à-dire en ajoutant une phase de calcul de cepstres par prédiction linéaire après le calcul des énergies des bandes critiques (taux d'erreur **10.7%**). Ce n'est malheureusement pas possible dans l'approche qui nous intéresse car nous devons garder des paramètres de type fréquentiel. En effet, nous souhaitons reconstruire les portions spectro-temporelles bruitées ou filtrées sur base des portions non-bruitées. Ils convient donc d'opérer la reconstruction dans les domaine spectral de façon à pouvoir isoler les bruits colorés ou les bandes de fréquence filtrées. Cette remarque montre cependant l'importance majeure d'une recherche concernant les paramètres acoustiques de type fréquentiel. Nous y reviendrons a la Section 4.

Ensuite, nous avons envisagé l'utilisation d'un système basé sur l'approche illustrée à la Figure 1. Plusieurs configurations ont été envisagées en ce qui concerne la distribution de probabilité des observations. Le module de détection des données manquantes est basé sur un estimateur de rapport signal/bruit utilisant des histogrammes d'énergie dans les différentes bandes critiques [2]. Les composantes avec un rapport signal/bruit local inférieur à 10 dB étaient considérées comme manquantes.

3.1. *Matrices de Covariance Diagonales*

Dans un premier temps, nous avons considéré des distributions multi-gaussiennes à matrice de covariance diagonale comprenant 1, 16 et 64 gaussiennes. Dans ce cas, la reconstruction est quasi immédiate car le deuxième terme de l'addition dans l'Equation 3 est nul. Cette approche revient donc à remplacer les données manquantes par leur moyenne, ce qui est très simple et, surtout, ne permet pas d'utiliser la corrélation entre les éléments du vecteur acoustique. Comme illustré à la Figure 2, cette approche n'apporte une amélioration significative des performances que dans des cas extrêmes de filtrage et pour une distribution comportant un nombre très important de gaussiennes. Dans les autres cas, la distorsion introduite par la reconstruction est vraisemblablement supérieure à celle introduite par la perturbation. Cette configuration a également été envisagée dans [6]. Les expériences effectuées indiquent également de faibles performances.

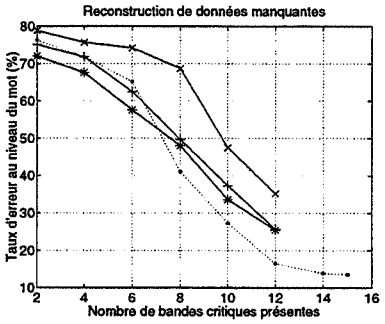


Figure 2: Taux d'erreurs au niveau du mot pour différentes conditions de filtrage passe-bas. La courbe en pointillés correspond au système utilisant l'ensemble des paramètres spectro-temporels. Les courbes continues correspondent à différentes configurations du système utilisant la détection automatique et la reconstruction des données manquantes. Les courbe '×', '+' et '∗' correspondent respectivement à des distributions modélisées par 1, 16 et 64 gaussiennes à **matrices de covariance diagonales**.

3.2. *Matrices de Covariance Pleines*

Des résultats plus encourageant ont été obtenus en utilisant des matrices de covariance pleines. L'emploi de matrices pleines est justifié par la forte corrélation caractérisant les différentes composantes fréquentielles du vecteur acoustique. Ces modèles exigent un grand nombre d'inversions de matrices, ce qui risque de devenir fort lourd. Les résultats obtenus (Figure 3) indiquent cependant une nette amélioration par rapport au système de base, avec seulement une gaussienne multidimensionnelle. L'emploi d'un nombre plus important de gaussiennes, 32 en l'occurence, apporte encore plus de robustesse dans le cas de faibles perturbations mais, curieusement, est sensiblement moins bon dans le cas de fortes perturbations (filtrage de plus de 50% du spectre).

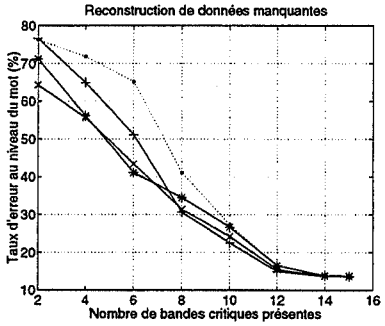


Figure 3: Les courbes continues correspondent au système utilisant la reconstruction sur base de distributions de probabilités gaussiennes à **matrices de covariance pleines**. Les courbes '×' et '+' correspondent respectivement à des distributions modélisées par 1 seule gaussienne et par 32 gaussiennes. La courbe '∗' correspond à une expérience additionnelle pour laquelle nous avons utilisé une connaissance a priori des portions spectrales manquantes et une reconstruction sur base d'une seule gaussienne.

Finalement, nous avons envisagé le cas d'un signal perturbé par un bruit additif sinusoïdal stationnaire. Les résultats sont présentés à la Figure 4. Pour comparaison, cette figure reprend également une courbe correspondant au système multibande proposé dans [4], présentant de très bonnes propriétés de robustesse aux bruits colorés. Nous constatons à la Figure 4 qu'à des niveaux de bruit très élevés, le système basé sur la reconstruction des données manquantes surpasse les deux autres systèmes. Les excellentes performance obtenues ici s'expliquent par le fait que seules deux bandes critiques (à cause du recouvrement des filtres implémentant les bandes critiques) sont corrompues et que leur reconstruction peut s'effectuer de façon très fiable sur base des autres composantes et leurs dérivées premières. Notons que les paramètres acoustiques utilisés ne sont pas les mêmes pour les différentes expériences. Dans le cas de l'approche basée sur la reconstruction, il s'agit de bandes critiques, alors que pour les autres systèmes, il s'agit de coefficients PLP. Cela justifie les différences de performances dans le cas de parole faiblement bruitée. Nous reviendrons sur ce problème à la section suivante.

Illustrons finalement cet article par des exemples de détection et de reconstruction des données manquantes (Figure 5).

4. DISCUSSION

L'approche proposée exige l'utilisation de paramètres de type spectraux. Nos expériences ont été effectuées avec les énergies de bandes critiques. Nous pourrions envisager

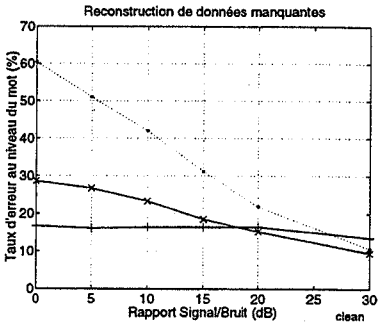


Figure 4: Taux d'erreurs au niveau du mot pour différents niveaux de bruits (ton pur à 400 Hz, rapport signal/bruit global de 0 dB à 30 dB). La courbe en pointillés correspond au système de référence. La courbe '×' correspond aux résultats obtenus avec l'approche multibande. La courbe '+' correspond au système utilisant la reconstruction des données manquantes sur base d'une distribution à une seule gaussienne à matrice de covariance pleine.

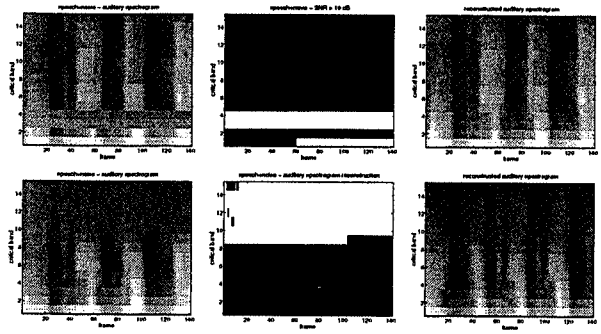


Figure 5: Détection et reconstruction des données manquantes pour un signal perturbé par un bruit additif sinusoïdal à 400 Hz avec un rapport S/B de 0 dB (haut) et filtré par un filtre passe-bas (bas). Les graphiques de gauche présentent le résultat de l'analyse spectro-temporelle du signal bruité. Les graphiques centraux présentent le résultat du détecteur de données manquantes. Les zones blanches correspondent aux portions manquantes. Les graphiques de droite correspondent à la version reconstruite des spectrogrammes, en utilisant une seule gaussienne à matrice de covariance pleine.

une phase de calcul de coefficients cepstraux sur base des vecteurs acoustiques reconstruits. Les cepstres conduisent généralement à de meilleures performances de reconnaissance que les paramètres spectraux. Des travaux sur l'approche multibande [4] ont montré que le calcul de cepstres par prédiction linéaire sur des bandes de fréquences étroites conduit à une amélioration des performances par rapport à l'utilisation directe des bandes critiques. Nos expériences [4] ainsi que d'autres recherches [9] confirment cette constatation pour différentes configurations de bandes Ce type de paramètres pourrait également être utilisé dans notre cas. Dans ces expériences, nous avons également constaté que l'augmentation du nombre de bandes de fréquence au delà de quatre, entraîne une dégradation des performances. Ce phénomène est vraisemblablement dû à l'hypothèse de non-corrélation qui est posée et qui est loin d'être vérifiée si les bandes de fréquences sont très étroites et rapprochées. Un avantage de l'approche présentée ici est que le problème de corrélation des bandes de fréquences ne se pose pas.

5. CONCLUSIONS

Cet article présente une technique de reconstruction de données manquantes et son utilisation dans le cadre de systèmes de reconnaissance automatique de parole utilisant des réseaux de neurones artificiels conjointement aux modèles de Markov cachés. Les résultats présentés montrent qu'un système de reconstruction relativement simple, piloté par un estimateur de rapport signal/bruit élémentaire, confère au système de reconnaissance une importante robustesse au filtrage ainsi qu'au bruit additif fortement coloré. Des expériences complémentaires sont nécessaires pour valider cette approche pour des bruits additifs réels, bruits qui peuvent couvrir une plage fréquentielle très importante. Dans ce cas, nous entrevoyons la possibilité de n'utiliser que les portions spectro-temporelles pour lesquelles l'énergie du signal de parole est bien supérieure à l'énergie du bruit (zones spectrales autour des formants et portions temporelles voisées).

BIBLIOGRAPHIE

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE ASSP*, vol. 2, no. 27, 1979.

[2] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards sub-band-based speech recognition," in *Proc. of European Signal Processing Conference*, (Trieste, Italy), pp. 1579–1582, Sept. 1996.

[3] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, ISBN 0-7923-9396-1, 1994.

[4] H. Bourlard and S. Dupont, "Sub-band-based speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), pp. 1251–1254, Apr. 1997.

[5] R. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at cslu," in *Proc. of Intl. Spoken Language Processing*, (Yokohama, Japan), September 1994.

[6] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), Apr. 1997.

[7] M. Gales, "Nice model-based compensation schemes for robust speech recognition," in *Proc. of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 55–64, Apr. 1997.

[8] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. of ICASSP'97*, (Munich), pp. 1255–1258, 1997.

Rehaussement par soustraction spectrale et compensation des paramètres manquants pour la reconnaissance robuste du locuteur et de la parole

Mounir El-Maliki *Philippe Renevey* *Andrzej Drygajlo*

Laboratoire de Traitement des Signaux
École Polytechnique Fédérale de Lausanne
1015 Lausanne
e-mail: (Mounir.Elmaliki, Philippe.Renevey, Andrzej.Drygajlo)@epfl.ch

ABSTRACT

In the framework of Gaussian mixture models (GMMs) and hidden Markov models (HMMs) we present a new approach towards robust automatic speaker verification and speech recognition in adverse conditions. This new and simple approach is based on the combination of the generalized spectral subtraction performed in critical bands for speech enhancement and a missing feature compensation in the GMMs and HMMs. The identity of spectral features missing due to noise masking is provided by the spectral subtraction algorithm. The statistical missing feature compensation modifies dynamically the probability computations performed in GMM and HMM recognizers. In this paper we show that this method, which integrates the process of noise masking directly in the framework of probabilistic modeling, improves noise compensation techniques used for robust speaker verification and speech recognition.

1. INTRODUCTION

Dans les applications réelles, les systèmes de reconnaissance automatique du locuteur et de la parole sont confrontées à des conditions adverses. Il est donc impératif de considérer le problème du bruit ambiant. L'apprentissage des systèmes se fait généralement sur des données de bonne qualité alors que la reconnaissance peut s'effectuer sur des données bruitées. Les différences entre les conditions d'apprentissage et de reconnaissance réduisent considérablement les performances. Le choix du type d'analyse et de modèles statistiques permet de réduire l'influence de l'environement. Des techniques permettant d'adapter les paramètres, ou les modèles aux conditions de bruits améliorent encore les résultats. Dans ce cadre, nous avons testé la soustraction spectrale généralisée afin de réduire la différence entre les conditions d'apprentissage et de reconnaissance.
Nous proposons une nouvelle technique d'adaptation au bruit basée sur la soustraction spectrale. Elle permet simultanément de rehausser le signal et de détecter les paramètres masqués par le bruit. Le classificateur tient dynamiquement compte de cette information pour éliminer ces paramètres (paramètres manquants) lors de la reconnaissance.
Contrairement à l'approche de rehaussement en tant qu'étape de prétraitement à un système de reconnaissance qui essaie d'estimer le signal propre, la méthode proposée permet d'éliminer l'influence des paramètres masqués par le bruit ambiant qui est l'une des causes principales de dégradation des performances de reconnaissance.
Cette méthode a été testée sur des systèmes de vérification du locuteur (mode indépendant du texte) basée sur les modèles à mélanges de gaussiennes (GMM) et de reconnaissance de la parole en utilisant les modèles de Markov cachés (HMM).

2. REHAUSSEMENT DE LA PAROLE PAR SOUSTRACTION SPECTRALE

Des techniques de rehaussement sont souvent employées dans des systèmes de reconnaissance du locuteur et de la parole afin d'en améliorer les performances [Ace93, Han94, LB92, OG96]. La soustraction spectrale a été développée afin de réduire l'influence d'un bruit additif dans le domaine spectral [Bol79]. Une de ses variantes est la soustraction spectrale de puissance [LO79]. Une valeur estimée du spectre de puissance du bruit est soustraite au spectre de puissance du signal bruité. La soustraction spectrale de puissance est définie par la relation suivante:

$$|\hat{S}_m(\omega)|^2 = \begin{cases} |Y_m(\omega)|^2 - |\overline{N}(\omega)|^2 & \text{si } |Y_m(\omega)|^2 > |\overline{N}(\omega)|^2, \\ 0 & \text{autrement.} \end{cases} \quad (1)$$

où $|Y_m(\omega)|^2$ est le spectre de puissance du signal bruité pour la fenêtre d'indice m et $|\overline{N}(\omega)|^2$ représente le spectre de puissance moyen du bruit, estimé pendant les pauses.
Un problème principal de cette technique est la non-linéarité liée à la mise à zéro de certains des coefficients du spectre de puissance. Dans le cas de rehaussement de la parole pour auditeurs humains, ces non-linéarités ajoutent au signal un bruit possèdant une nature périodique (bruit musical). Pour réduire l'effet dérangeant de ce bruit, une soustraction spectrale généralisée (SSG) à été proposée [BSM79]. Elle peut être exprimée de la manière suivante:

$$D_m(\omega) = |Y_m(\omega)|^2 - \alpha \cdot |\overline{N}(\omega)|^2 \quad (2)$$

et

$$|\hat{S}_m(\omega)|^2 = \begin{cases} D_m(\omega) & \text{si } D_m(\omega) > \beta \cdot |\overline{N}(\omega)|^2, \\ \beta \cdot |\overline{N}(\omega)|^2 & \text{autrement.} \end{cases} \quad (3)$$

avec $\alpha \geq 1$ et $0 < \beta \leq 1$
Le rôle du facteur α est de surestimer le spectre du bruit, forçant en cela les pics spectraux du bruit résiduel à être plus atténués. Cela premet d'améliorer la réduction du

bruit résiduel, mais par contre, la distorsion du signal s'en trouve augmentée. Le facteur β permet de fixer la valeur minimale en dessous de laquelle le spectre de puissance du signal propre estimé ne peut descendre. Cette valeur minimale permet de remplir les vallées dans le domaine spectral et ainsi de réduire l'influence du bruit résiduel par rapport au cas où $\beta = 0$. Dans ce papier, la soustraction spectrale généralisée n'est pas utilisée uniquement pour le rehaussement de la parole, mais aussi afin de détecter trame par trame les bandes de fréquences dans lesquelles le bruit masque le signal de parole. Ces bandes ne seront pas utilisées pour le calcul du score de reconnaissance.
Sous l'hypothèse que le bruit est additif et stationnaire, la détection des bandes de fréquences qui ne comporte pas une information pertinante relative au signal de parole peut être effectuée en utilisant la soustraction spectrale de puissance généralisée. Les bandes de fréquence dont la puissance du signal réhaussé $D_m(\omega) = |Y_m(\omega)|^2 - \alpha \cdot |\overline{N}(\omega)|^2$ descend en dessous de la valeur fixé par $\beta \cdot |\overline{N}(\omega)|^2$ sont considérées comme des paramètres manquants pour la reconnaissance. Cette détection peut s'effectuer dans diverses conditions de bruits. Le système de reconnaissance est ensuite modifié dynamiquement afin de ne pas tenir compte des paramètres manquants.

3. COMPENSATION DES PARAMÈTRES MANQUANTS DANS LE MODÈLE PROBABILISTE

L'utilisation d'une détection automatique de paramètres manquants dans des bandes critiques permet d'inclure une information au système de reconnaisance sur la pertinance des sous-bandes. En effet, l'estimation du signal propre par l'algorithme de la soustration spectrale est incomplète dans les bandes où le bruit masque le signal de la parole. Le besoin d'une estimation "erronée" de ces bandes masquées se trouve injustifié car la reconstruction du signal de la parole n'est pas nécessaire vu que le rehaussement n'est pas déstiné à une oreille humaine. Dans ce cas, la compensation des paramètres manquants se fera par leur élimination du processus de reconnaissance automatique.
Les sytèmes de reconnaissance probabilistes tels que les HMMs et les GMMs permettent de faire la compensation sans toucher au coût du calcul. Pour plus de clareté, on considère dans les formules suivantes une chaîne de Markov cachée multi-gaussiennes à un seul état. Pour le vecteur de paramètres $\mathbf{x}$, la fonction de la densité de probabilité (fdp) s'exprime sous forme d'une combinaison pondérée des fdps gaussiennes sous la forme suivante:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \Phi_i(\mathbf{x}, \mu_i, \Sigma_i) \qquad (4)$$

où p_i est le poids associé à la fdp gaussienne Φ_i, Φ_i est une fdp de dimension D définie par le vecteur de moyennes μ_i et la matrice de covariance Σ_i et M est le nombre total des fdps gaussiennes. On suppose que l'apprentissage des HMMs se fait sur des données relativement propres afin de constituer une réference conforme à la distribution réelle du signal de la parole. Ainsi, la seule modification à apporter au classificateur se situe dans la phase de reconnaissance.

Dans le cas d'une matrice de covariance diagonale[1], l'Eq. 4 se transforme en un produit de fdp gaussiennes mono-dimensionnelles selon:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2) \qquad (5)$$

où μ_{ji} est la moyenne et σ_{ji}^2 est la variance de la composante j du vecteur de paramètres $\mathbf{x}$. Chaque vecteur de paramètres $\mathbf{x}$ de la séquence $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ issue de la parole bruitée destinée à la reconnaissance se compose en deux sous-vecteurs $\mathbf{x}_p$ et $\mathbf{x}_m$. Les sous-vecteurs $\mathbf{x}_p$ et $\mathbf{x}_m$ représentent les composantes du vecteur de paramètres présents (utiles) et manquants, respectivement. Eq. 5 s'exprime comme suit:

$$p(\mathbf{x}|\lambda) = \qquad (6)$$
$$\sum_{i=1}^{M} p_i \prod_{\substack{j \\ present}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2) \cdot \prod_{\substack{j \\ manquant}} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2)$$

La fonction de vraissemblance partielle calculée sur la partie présente des paramètres peut être déduite en éliminant le deuxième terme du produit de l'équation précédente [LC97, CGAA94]:

$$p(\mathbf{x_p}|\lambda) = \sum_{i=1}^{M} p_i \prod_{\substack{j \\ present}} \Phi_i(x_j, m_{ji}, \sigma_{ji}^2) \qquad (7)$$

Le schéma global du système de reconnaissance est illustré dans la figure 1. La soustraction spectrale est effectuée dans les bandes organisées selon une échelle Bark. Le signal bruité est rehaussé et en même temps les paramètres manquants sont détectés pour ensuite être dynamiquement éliminés du processus de reconnaissance par le classificateur probabiliste.

4. VÉRIFICATION DU LOCUTEUR INDÉPENDANTE DU TEXTE PAR LES GMMS

4.1. *Expériences et résultats*

L'approche présentée dans cet article est appliquée dans le cadre de la vérification du locuteur indépendante du texte par la modèlisation GMM [Rey95]. Un sous-ensemble de locuteurs constitué de 22 personnes est sélectionné de la base de données NTIMIT de qualité téléphonique. 14 valeurs d'énergie des bandes critiques sont calculées et la soustraction spectrale généralisée est utilisée pour le rehaussement et la sélection des paramètres manquants. Environ 24 secondes de parole sont utilisées pour l'apprentissage du modèle du locuteur et deux phrases de 3 secondes chacune sont déstinées au test. Pour générer des signaux bruités, un bruit blanc gaussien artificiel et un bruit de cockpit d'avion F16 pris de la base NOISEX-92

[1] L'indépendance statistique entre les composantes du vecteur de paramètres n'est pas assurée. Le choix de la matrice de covariance diagonale n'est justifié que pour des raisons de réduction du temps de calcul dû à la compensation des paramètres manquants (CPM). Des détails sur cette compensation avec une matrice de covariance pleine se trouvent dans [CMG97]

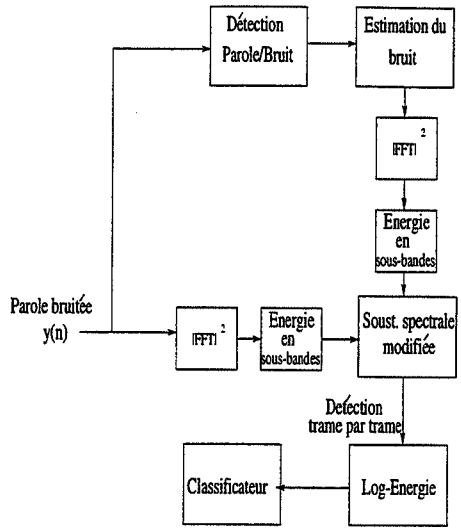


Figure 1: Le schéma global de la reconnaissance avec une compensation des paramètres manquants

ont été ajoutés aux données de test en utilisant différents rapports signal-sur-bruit (SNR). Au cours des premières expériences, le facteur α d'une valeur de 3 a donné la meilleure mesure de l'EER (Equal Error Rate).

Table 1: *La mesure de l'EER dans le cas d'un bruit de cockpit d'avion.*

Experiences (SNR)	EER[%] (9dB)	EER[%] (12dB)	EER[%] (15dB)
parole bruitée	37.2	32.3	28
SSG	28.2	25.3	23.1
CPM	**16.2**	**15.7**	**12.7**

Table 2: *La mesure de l'EER dans le cas d'un bruit blanc gaussien.*

Experiences (SNR)	EER[%] (9dB)	EER[%] (12dB)	EER[%] (15dB)
parole bruitée	43.6	37.5	34.8
SSG	35.8	32.6	28.1
CPM	**20.2**	**18.9**	**15.1**

Les tests effectués sur la base de données sans bruit ont donné une mesure de l'EER égale à 11.5%. D'après les tableaux 1 et 2, le système de vérification du locuteur en présence de bruit subit une forte dégradation de performances. La méthode de la soustraction spectrale généralisée employée pour le rehaussement de la parole permet de contribuer à l'amélioration du taux de reconnaissance. On peut encore constater que l'approche proposée de détection et de compensation de paramètres manquants (CPM) permet de réduire fortement l'EER dans tous les tests.

5. RECONNAISSANCE DE LA PAROLE PAR HMMS

5.1. *Système utilisé pour la reconnaissance de la parole*

Le système de reconnaissance de la parole est basé sur les modèles de Markov cachés (HMMs). Les paramètres utilisés sont le logarithme de l'énergie des sorties d'un banc de filtres dont les bandes sont réparties selon une échelle de Mel entre 0 et 8 kHz ainsi que l'énergie totale du bloc et les dérivées premières et secondes de ces paramètres (au total 75 paramètres: 24 paramètres spectraux + 1 énergie + 25 dérivées premières + 25 dérivées secondes). Les fenêtres d'analyse ont une durée de 25 ms et un recouvrement de 15 ms. Le système de reconnaissance est basé sur une reconnaissance par phonèmes. Ces modèles ont été entrainés sur la base de données TIMIT et les tests de reconnaissance ont été effectués sur les chiffres en anglais de la base de données TIDIGIT.
Le système permettant la détection des segments de parole doit être aussi précis que possible. Il doit pouvoir fonctionner de manière satisfaisante en présence de bruit car la qualité de l'estimation du bruit dépend de la qualité de sa segmentation. La méthode utilisée se base sur un seuillage de l'entropie du spectre du signal de parole [AMB97].
La valeur moyenne des paramètres du bruit est ensuite calculée sur les segments de "silences" détectés en prenant la moyenne arithmétique dans chacune des bandes de fréquences.
Les différentes bandes de fréquences sont compensées en utilisant Eq. 3 avec $\alpha = 1$ et $\beta = 0$. Les bandes de fréquence dont la valeur est égale à zéro après la soustraction spectrale sont considérées comme manquantes et ne seront pas utilisées dans la reconnaissance.
Le système de reconnaissance utilise ensuite Eq. 7 pour calculer les scores des modèles. Les paramètres dynamiques (dérivées premières et secondes des valeurs des paramètres) ne sont pas compensés et sont considérés comme n'étant pas touchés par le bruit.

5.2. *Résultats expérimentaux*

Les résultats de reconnaissance ont été obtenus sur une partie de la base de données TIDIGIT qui contient des chiffres en anglais. Les phrases contenant un seul chiffre ont été utilisées. Les paramètres des bancs de filtres (BF) et "Mel Frequency Cepstral Coefficients" (MFCC) servent de référence de comparaison pour mesurer les performances de la méthode proposée CPM. Le bruit additionné est un bruit blanc filtré passe bande par un filtre RIF. Le SNR détermine le rapport entre l'énergie moyenne du signal pendant les tranches de parole et celle du bruit. Le logiciel HTK a été utilisé pour l'apprentissage et la reconnaissance. La méthode proposée a été programée en modifiant ce dernier.

5.3. *Discussion des résultats*

Les résultats de reconnaissance présentés dans le tableau 3 montrent que la méthode proposée améliore de façon significative les résultats de reconnaissance pour des bruits

Table 3: Taux de reconnaissance pour différents types de bruit et SNR

SNR	6dB	9dB	12dB	15dB
Param.	% correct			
Bruit de bande 1-2 kHz				
MFCC	9.09	9.09	11.36	20.45
BF	27.27	27.27	45.45	56.82
CPM	**34.09**	**45.45**	**63.04**	**72.73**
Bruit de bande 1-3 kHz				
MFCC	18.18	29.55	34.09	36.36
BF	27.27	34.09	45.45	52.27
CPM	**36.36**	**45.45**	**56.82**	**65.91**
Bruit de bande 2-3 kHz				
MFCC	25.00	29.55	34.09	36.36
BF	29.55	36.36	40.91	59.09
CPM	**40.91**	**50.00**	**68.18**	**72.73**
Bruit de bande 5-7 kHz				
MFCC	**34.09**	**34.09**	**40.91**	**54.55**
BF	9.09	9.09	15.91	20.45
CPM	15.91	20.45	27.27	45.45
Bruit de bande 6-7 kHz				
MFCC	**50.00**	**54.55**	**59.09**	63.64
BF	11.36	15.91	25.00	48.84
CPM	22.73	34.09	52.27	**68.18**

de bandes comprises entre 1 et 3 kHz. Pour les bruits de bande situés entre 5 et 7 kHz, les résultats sont moins bons que ceux obtenus par les MFCCs. Le type de système de reconnaissance choisi pour les tests explique cette différence. En effet, le système est basé sur la reconnaissance des chiffres en anglais. Beaucoup de sons fricatifs sont présents dans ces chiffres ("three", "four", "five", "six", "seven" et "zero"), or la bande de fréquences 5 à 7 kHz contient une partie importante de l'énergie des sons fricatifs, et l'absence de ces bandes dans le système de reconnaissance diminue la capacité à pouvoir dicerner ces chiffres les uns des autres. Il faut cependant remarquer que la méthode proposée obtient systématiquement des meilleurs résultats que ceux obtenus en utilisant uniquement les paramètres des bancs de filtres.

6. CONCLUSION

Dans cet article, nous avons montré que la compensation dynamique des paramètres manquants dans le processus de reconnaissance permet d'augmenter les performances dans des conditions adverses, ceci aussi bien dans les GMMs que dans les HMMs. La méthode proposée est basée sur la soustraction spectrale généralisée. Elle permet simultanément de rehausser le signal de parole et de détecter les paramètres manquants. Les résultats obtenus dans diverses conditions de bruit montrent clairement les bonnes performances obtenues par cette méthode pour la reconnaissance du locuteur et de la parole.

BIBLIOGRAPHIE

[Ace93] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, 1993.

[AMB97] I. Abdallah, S. Montrésor, and M. Baudry. Speech signal detection in noisy environment using a local entropic criterion. In *Proc. EUROSPEECH'97*, volume 5, pages 2595–2599, Rhodes, Sep. 1997.

[Bol79] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 27(2):113–120, April 1979.

[BSM79] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of Speech Corrupted by Acoustic Noise. In *Proc. of ICASSP'79*, pages 208–211, April 1979.

[CGAA94] M. Cooke, P.D. Green, C. Anderson, and D. Abberley. Recognition of occluded speech by hidden Markov models. Technical Report TR-94-05-01, 1994.

[CMG97] M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *Proc. ICASSP*, pages 863–866, Munich, April 1997.

[Han94] J.H.L Hansen. Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect. *IEEE Trans. on Speech and Audio Proc.*, 2(4):598–614, Oct. 1994.

[LB92] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtraction (NSS) and hidden Markov models and projection for robust recognition in cars. *Speech Communications*, 11(2-3):215–228, June 1992.

[LC97] R. P. Lippman and B. A. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Proc. EUROSPEECH'97*, volume 1, pages KN 37–40, Rhodes, Sep. 1997.

[LO79] J.S. Lim and A.V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. of the IEEE*, 67(12):1586–1604, 1979.

[OG96] J. Ortega-García and J. Gonzàlez-Rodríguez. Overview of speech enhancement techniques for automatic speaker recognition. In *Proc. ICSLP*, volume 2, pages 929–932, Oct. 1996.

[Rey95] D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

La détection de nouveauté basée sur le temps de stabilisation d'un réseau de neurones: application possible en reconnaissance de parole?

Jean ROUAT[1], *Tuong Vinh HO*[1,2]

[1] ERMETIS, DSA, Université du Québec à Chicoutimi
555 boul. de l'Université, CHICOUTIMI, Québec, CANADA, G7H 2B1
Tél.: 1 418 545 5011 x5642 - Fax: 1 418 545 5012
[2] Ecole Polytechnique de Montréal, Canada
e-mail: Jean_Rouat@uqac.uquebec.ca, vho@uqac.uquebec.ca

ABSTRACT

We propose a spiking neural network model inspired from a simulated cortex model. Also, a new paradigm for pattern recognition by neural networks with complex dynamics is presented. The 'relaxation' time of the network is used as a criterion for novelty detection. We compare the proposed neural network with Hopfield and backpropagation networks for a noisy digit recognition task. It is shown that the proposed network is more robust. We also design a limited experiment based on the recognition of temporal sequences of vowels and we show that the network is able to perform the recognition with a rate of 100% (sequences of 5 and 11 vowels). Regarding speech and pattern recognition tasks, the proposed spiking network seems to have a strong potential.

1. INTRODUCTION

Il existe des systèmes de reconnaissance de parole indépendants du locuteur pour des vocabulaires limités et qui offrent des performances intéressantes à condition que l'environnement ne soit pas trop corrompu par le bruit et les interférences [Dup97]. Toutefois, l'interrogation de bases de données à distance, (téléphone, radio, etc.), la transcription des nouvelles télévisées ou radio par exemple, nécessitent de concevoir des systèmes polyvalents et peu sensibles aux conditions et aux environnements d'opérations ainsi qu'aux bruits ambiants (téléphone cellulaire, cabine téléphonique, voitures, avions, camions, etc.)[Ezz97][Rob97]. Or le traitement de la parole en milieu bruyant est loin d'être résolu et les systèmes de reconnaissance actuels ne peuvent fonctionner dans de tels environnements sans observer de dégradation significative des performances [Gau97][Woo97][Bak97].

La mise en place de système de reconnaissance versatile fait souvent appel à des analyses perceptives inspirées de modèles du système auditif [Pat95]. Ces analyses codent l'information liée à la parole dans un espace à représentation spatio-temporelle complexe et difficile à caractériser via les algorithmes couramment utilisés en reconnaissance de parole. En effet, ces représentations spatio-temporelles sont structurées à la fois spectralement et dans le temps. L'information est en général non stationnaire. Il y a donc lieu de mettre au point des paramètres et des algorithmes de reconnaissance de formes capables d'exploiter cette information temporelle. Or une majorité des algorithmes contemporains assument une certaine stationnarité du signal et des paramètres d'analyse. De plus, ces systèmes requièrent un apprentissage supervisé souvent long et fastidieux. Nous pensons qu'il y a lieu d'étudier l'élaboration de nouvelles techniques d'analyse et de reconnaissance adaptées à l'exploitation de l'information temporelle fine telle que générée par exemple via des analyses d'inspiration perceptive. Une information temporelle fine serait par exemple celle qui est liée à l'enveloppe des signaux à la sortie d'un banc de filtres cochléaires [Rou97].

2. MOTIVATION

Il apparaît intéressant d'élaborer de nouveaux outils aptes à traiter une information dynamique et non stationnaire. Ce type d'outil devrait pouvoir être utilisé en reconnaissance de formes afin de traiter une information codée dans le temps (parole, par exemple). Nous proposons un système qui devrait être en mesure de remplir ces critères.

Il s'agit d'un système de détection de la nouveauté basé sur le temps de stabilisation d'un réseau de neurones (dit bio-inspiré) dont l'architecture est inspirée de la couche IV du cortex. L'originalité du travail réside principalement dans la définition d'une règle de modification des connexions synaptiques et dans la création du critère de temps de stabilisation pour la reconnaisance. En effet, la majorité des techniques de reconnaissance des formes utilisant les réseaux de neurones (souvent formels) se base sur des critères relativement statiques (minimisation d'une fonction d'erreur, maximisation d'une probabilité, etc.) et codent l'information temporelle de façon statique via la structure spatiale des entrées [Hay94]. Par ailleurs, l'approche adoptée ici ne nécessite pas de supervision du réseau. Celui-ci est en mesure de détecter la nouveauté et de s'adapter de façon autonome. De plus, il n'y a pas de différence entre apprentissage et reconnaissance.

Dans un premier temps, il est important de tester et de valider ce type d'approche vis-à-vis des systèmes plus classiques afin de s'assurer que le réseau proposé puisse réaliser au minimum des tâches similaires à celles exécutées par les réseaux dits formels . Nous présentons donc une série de tests préliminaires effectués en reconnaissance de chiffres bruités ainsi qu'en reconnaissance de séquences de voyelles puis nous discuterons du potentiel de ce travail en reconnaissance de parole.

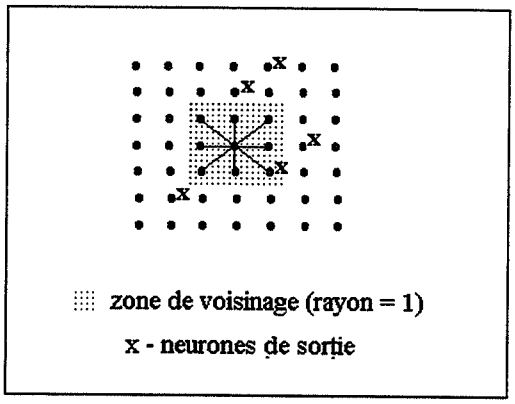


Figure 1: Architecture du réseau de neurone.

3. LE RÉSEAU DE NEURONES

3.1. *Modèle du neurone*

Le neurone est de type intégration et décharge ('Integrate and Fire') avec période réfractaire et potentiel post-synaptique décroissant. L'état du neurone à l'instant t est caractérisé par son potentiel interne U. La réponse $S_i(t)$ du neurone i est donnée par:

$$S_i(t) = \begin{cases} 0 & \text{pour } (t - t_{spike}) < \rho, \\ \mathcal{H}[U_i(t) - \theta] & \text{pour } (t - t_{spike}) \geq \rho \end{cases} \quad (1)$$

$\mathcal{H}$ est la fonction d'Heaviside. Le potentiel de contrôle $U_i(t+1)$ pour la cellule i est l'intégration de toutes les réponses afférentes à l'instant t et tient compte de la fréquence instantanée f_i de décharge de la cellule i.

$$U_i(t+1) = \sum_j C_{ij}(t)S_j(t) + U_i(t) + e_i(t) + f_i(t) \quad (2)$$

$C_{ij}(t)$ est la connexion de la cellule j vers la cellule i à l'instant t. $e_i(t)$ est le signal externe (entrée du stimulus) pour la cellule i considérée.

3.2. *Architecture et apprentissage*

L'architecture est inspirée du modèle de la couche IV du cortex tel que proposé par Hill & Villa [Hil97]. Ils définissent une couche à 2 dimensions comprenant des neurones inhibiteurs et excitateurs avec récursivité. Chaque neurone est interconnecté à ceux qui appartiennent à son ensemble de voisinage. Cet ensemble est défini comme étant carré, centré autour du neurone et comprend une taille différente selon le type de connexion (inhibitrice ou excitatrice). Nous choisissons les neurones de sortie du réseau de façon aléatoire à partir de neurones de type excitateurs (figure 1). Le signal de sortie est caractérisé par l'état de décharge de l'ensemble des neurones de sortie.

L'apprentissage est inspiré du travail de Stassinopoulos et Bak [Sta94]. Pour chaque signal d'entrée, l'action du réseau est considérée comme étant un succès si au moins un neurone appartenant à l'ensemble des neurones de sortie se décharge. Dans ce cas, les connexions entre les neurones actifs sont renforcées. Si l'action est sans succès, les connexions entre les neurones actifs sont affaiblies. La mise à jour des connexions se fait selon la règle d'Hebb. Soit h, le signal de feedback (h = +1 pour le renforcement ou -1 pour affaiblir). La mise à jour des poids entre deux neurones est donnée par:

$$C_{ij}(t+1) = C_{ij}(t) + \alpha C_{ij}(1 - C_{ij})S_i(t)S_j(t)h(t) \quad (3)$$

avec α le taux d'apprentissage.

Table 1: Taux d'erreurs pour un apprentissage sur les chiffres propres [0-4] et reconnaissance sur les chiffres [0-9] propres et bruités

Type de réseau	proposé	Hopfield	multicouches
Taux d'erreur (%)	10	21	27

Table 2: Taux d'erreurs pour un apprentissage sur les chiffres propres [5-9] et reconnaissance sur les chiffres [0-9] propres et bruités

Type de réseau	proposé	Hopfield	Multicouches
Taux d'erreur (%)	30	40	30

3.3. *Critère de reconnaissance*

Il n'y a pas de différence entre apprentissage et reconnaissance. La mise à jour des connexions synaptiques est toujours en cours sauf lorsque le réseau est stable. Après présentation d'un stimulus, le réseau a un comportement complexe qui finit par se stabiliser au cours du temps. Lorsque les changements de poids sont inférieurs à un seuil préétabli, on considère que le réseau est stable et que l'apprentissage ou la reconnaissance sont terminés. Le temps nécessaire T pour atteindre cet état stable est utilisé comme critère de détection de nouveauté. Ce temps T permet de caractériser le signal d'entrée. Un temps T très court (de l'ordre de 11 itérations) implique que le réseau a probablement déjà 'vu' ce signal au préalable. Cette notion permet de créer un système de détection de la nouveauté ayant un bon degré de robustesse vis-à-vis du bruit.

4. APPLICATION À LA RECONNAISSANCE DE CHIFFRES BRUITÉS

Des expériences de reconnaissance ont été réalisées à partir des chiffres 0 à 9 codés sur une matrice binaire de 7 par 5. Le réseau apprend uniquement sur les données non bruitées de 0 à 4 ou de 5 à 9 (colonne 1 ou 4 de la figure 2). Les expériences de reconnaissance portent sur les versions bruitées et propres des chiffres de 0 à 9 (toutes les colonnes de la figure 2). Le bruit est de type uniforme avec 20% des pixels modifiés. Le réseau est composé pour 70% de cellules excitatrices et pour 30% de

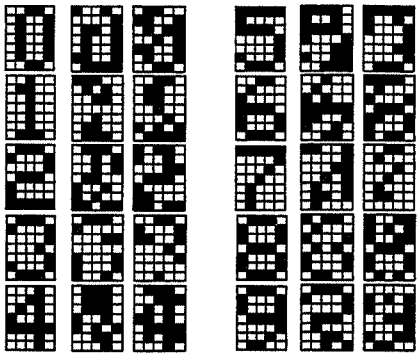


Figure 2: Images des chiffres propres (colonnes 1 et 4) et bruités (20% de bruit, colonnes 2,3,5 et 6).

cellules inhibitrices. Il comprend 49 cellules (7x7). Les chiffres sont codés sur une matrice de pixels 7x5. Chaque pixel est présenté sur l'entrée e_i d'une cellule choisie au hasard (i.e. 35 cellules sur les 49). Les tables 1 et 2 permettent de comparer le réseau proposé avec un réseau de Hopfield [Dem96] et un réseau multicouches utilisant la règle d'apprentissage généralisée avec moments [Ebe92]. Après un apprentissage à partir des chiffres propres de 0 à 4 (table 1) ou des chiffres propres de 5 à 9 (table 2), on présente l'ensemble des chiffres tel qu'illustrés à la figure 2. Le système doit ensuite différencier les chiffres qu'il connait (ceux qui étaient représentés dans la séquence d'apprentissage) de ceux pour lesquels il n'a jamais 'vu' de représentant. Pour la série apprise à partir de [0-4], le réseau proposé est nettement supérieur (table 1). Pour la série apprise à partir de [5-9], les performances sont similaires à celles du réseau multicouches.

5. RECONNAISSANCE DE SÉQUENCES DE VOYELLES

Le but de ces expériences est d'illustrer la possibilité d'identifier des séquences de voyelles déjà apprises et de les isoler des autres séquences pour lesquelles on retrouve les mêmes voyelles dans un ordre différent.

5.1. Les données de parole et la taille du réseau

Les données de parole sont disponibles sur le site de [Mer98]. Nous utilisons les voyelles prononcées par un seul locuteur pour onze mots anglais: heed(i), hid(I), head(E), had (A), hard (a:) hud (Y), hod(O), hoard(C:), hood (U), who'd(u:) et heard(3:). Le signal est filtré passe-bas à 4.7 kHz puis il est échantillonné à 10 kHz et quantifié sur 12 bits. Une analyse LPC d'ordre 12 est réalisée afin d'extraire 10 coefficients par fenêtre de signal. Une fenêtre de Hamming de 512 points est centrée au préalable sur la zone stable de chaque voyelle. On utilise donc 10 paramètres LPC par voyelle. Le réseau comprend 10 cellules (matrice de 5x2). Chaque paramètre est présenté sur l'entrée e_i d'une cellule.

5.2. Méthodologie

Une première séquence de voyelles est utilisée comme séquence d'apprentissage. Chaque voyelle de cette séquence est présentée au réseau de façon séquentielle pendant un intervalle de temps choisi T. Par exemple, la première voyelle est présentée au réseau pendant T itérations. Ensuite, la deuxième voyelle est présentée pour la même durée T. Il n'y a pas d'initialisation du réseau entre les présentation successives. On procède ainsi pour l'ensemble des voyelles de la séquence.

Après apprentissage, le comportement du réseau permet de savoir si celui-ci a identifié la séquence déjà apprise parmi les séquences qui lui sont présentées. Nous utilisons le temps de relaxation du réseau comme étant caractéristique du comportement. Chaque voyelle est présentée de façon séquentielle au réseau. Lorsque la première voyelle est présentée, le réseau oscille et se dirige vers un état d'équilibre. Ensuite, la deuxième voyelle est présentée, le réseau oscille à nouveau et atteint un autre état d'équilibre. On procède ainsi pour toutes les voyelles de la séquence. Après présentation de la séquence de voyelles, on obtient une série de temps de relaxation correspondant à la séquence de voyelles. En d'autres termes, nous avons effectué une 'projection' de la séquence de voyelles sur une séquence de temps de relaxation tout en réduisant la dimension des paramètres. L'analyse de cette séquence de temps permet d'indiquer si la série de voyelles a déjà été vue par le réseau.

Afin de faciliter l'analyse des séquences de temps, nous utilisons une méthode de codage très simple. Chaque intervalle de temps est codé par un chiffre. Par exemple, si le temps T se situe dans l'intervalle de 0 à 100 itérations ($0 < T \leq 100$), le chiffre associé est 0. De même, si le temps T se situe dans l'intervalle de 100 à 200 itérations ($100 < T \leq 200$), le code associé sera de 1, etc. De cette façon, on peut coder une séquence de temps de relaxation en une séquence de nombres entiers. En comparant les séquences de nombres correspondant aux séquences de voyelles de test avec la séquence déjà apprise, on peut savoir si une séquence de voyelles a déjà été 'vue' par le réseau. Les expériences suivantes illustrent ce principe.

5.3. Expérience 1: reconnaissance de séquences de 5 voyelles

Une séquence de 5 voyelles des mots anglais heed(i), hid(I), head(E), had (A) et hard (a:) est utilisée comme séquence d'apprentissage. Lors de l'apprentissage, chaque voyelle est présentée au réseau pour une durée de 200 itérations. Ensuite, cette même séquence est présentée au réseau afin de connaître le comportement de celui-ci face aux données déjà 'vues'. La séquence de temps de relaxation du réseau est de: 461, 11, 11, 245 et 145 itérations. En appliquant le codage précédent, on obtient la séquence de nombres suivante: 4, 0, 0, 2, 1. Pour les tests, nous utilisons un ensemble de séries de voyelles qui est composé de toutes les combinaisons possibles (120 combinaisons) des 5 voyelles déjà présentées au réseau lors de l'apprentissage. A titre d'exemple, la séquence de chiffres obtenue pour la série de voyelles (had (A), hid(I), head(E), heed(i), et hard (a:)) est 0, 0, 0, 5, 1. En comparant cette séquence à celle de l'apprentissage on peut conclure qu'elle n'avait jamais été vue par le réseau. Le taux d'erreur de reconnaissance des séquences est de 0% (au-

cune séquence en erreur).

5.4. *Expérience 2: reconnaissance de séquences de 11 voyelles*

Une séquence de 11 voyelles a été obtenue à partir des mots anglais heed(i), hid(I), head(E), had (A), hard (a:) hud (Y), hod(O), hoard(C:), hood (U), who'd(u:) et heard(3:)). Elle est utilisée comme séquence d'apprentissage. Chaque voyelle est présentée au réseau pendant une durée de 200 itérations. Lors des tests, nous utilisons un ensemble de séries de voyelles qui comprend 100 combinaisons des 11 voyelles d'apprentissage. Le taux d'erreur est de 0% (aucune séquence en erreur).

6. DISCUSSION ET CONCLUSION

Les expériences de reconnaissance de données statiques (les chiffres corrompus) indiquent que le réseau est en mesure de faire un travail aussi bon et souvent meilleur qu'un réseau de Hopfield et un réseau à rétropropagation.

Les expériences portant sur les séquences de voyelles permettent de mettre en valeur l'aptitude du réseau à traiter des séquences. En effet, la réponse du réseau (temps de stabilisation) à une voyelle dépend de ses états précédents. Pour une même voyelle, les caractéristiques dynamiques sont liées aux présentations antérieures.

Les expériences présentées sont relativement limitées et il y a lieu d'approfondir le travail afin de bien connaître les limites du réseau et son potentiel en reconnaissance de parole. Toutefois, sa robustesse au bruit ainsi que son aptitude à traiter de l'information structurée dans le temps, laissent à penser que ce système est un bon candidat pour la mise au point de techniques de reconnaissance de parole.

REMERCIEMENTS

Ce travail a été financé par le Conseil National de la Recherche en Sciences Naturelles et en Génie du Canada (CRSNG) ainsi que par la fondation de l'Université du Québec à Chicoutimi. Un merci tout particulier à Alessandro Villa pour les discussions stimulantes et enrichissantes en regard de ce travail.

BIBLIOGRAPHIE

[Bak97] R. Bakis, S.Schen, P. Gopalakrishnan, R. Gopinath, S. Maes and L. Polymenakos (1997). Transcription of broadcast news - system robustness issues and adaptation techniques. *Proc. of the IEEE-ICASSP*, Vol. 2, 711-714.

[Dem96] Demuth H., Beale M. (1996). Neural Network Toolbox for Use with MATLAB, *The Math Works Inc.*

[Dup97] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine and J.M. Boite (1997). Hybrid HMM/ANN systems for training independant tasks: Experiments on Phonebook and related improvements.*Proc. of the IEEE-ICASSP*, Vol. 3, 1767-1770.

[Ebe92] R. C. Eberhart, R. W. Dobbins. (1992). Neural Networks PC Tools: a practical guide. Academic Press. San Diego, USA.

[Ezz97] H. Ezzaidi, I. Bourmeyster and J. Rouat (1997). A new algorithm for double talk detection and separation in the context of digital mobile radio telephone. *Proc. of the IEEE-ICASSP*, Vol. 3, 1897-1900.

[Gau97] J.L Gauvain, G. Adda, L. Lamel and M. Adda-Decker (1997). Transcribing Broadcast News Shows.*Proc. of the IEEE-ICASSP*, Vol. 2, 715-718.

[Hay94] S. Haykin (1994). Neural Networks – A Comprehensive Foundation. *IEEE Computer Society Press* and *Macmillan College Publishing Company, Inc.*, 1994.

[Hil97] Hill S., Villa A. (1997). Dynamic transitions in global network activity influenced by the balance of excitation and inhibition. *Network: Computation in Neural Systems*, UK, Vol. 8, 2, 165-184.

[Ho97] Ho T.V., Rouat J. (1997). A Novelty Detector Using a Network of Integrate and Fire Neurons. *7th Int. Conf. on Artificial Neural Networks*, Lausanne, Switzerland, 8-10 Oct. 1997, Lecture Note on Computer Science (1327), Springer, pp. 103-108.

[Ho98] Ho T.V., Rouat J. (1998). Novelty Detection Based on Relaxation Time of a Network of Integrate-and-Fire Neurons. *International Joint Conference on Neural Networks*, Alaska, May 1998.

[Mer98] Merz, C.J., Murphy, P.M. (1998). UCI Repository of machine learning databases. *[http://www.ics.uci.edu/mlearn/MLRepository.html].* Irvine, CA: Univ. of California, Dep. of Information and Computer Science.

[Pat95] R.D. Patterson and M.H. Allerhand (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *J. Acoust. Soc. Amer.*, Vol. 98 (4), pp.1890-1894.

[Rob97] Robust speech recognition for unknown communication channels. *ESCA-NATO Tutorial and Research Workshop*, Pont-à-Mousson, France, 17-18 avril, 1997, éditions ESCA.

[Rou97] Spatio-temporal Pattern Recognition with Neural Networks: Application to Speech. *Artificial Neural Networks-ICANN'97, Lecture Notes in Comp.Sci.*,1327, Springer, 43-48.

[Rub95] A. J. Rubio and J. M. López (eds.) (1995). Speech Recognition and Coding, New Advances and Trends. *NATO ASI Series*, Springer.

[Sta94] Stassinopoulos D., Bak P. (1994). Self-organization in a Simple Brain Model. *Proc. of WCNN'94*, San Diego, Jun, Vol. 1, pp. 4-26.

[Woo97] P. C. Woodland, M. J.F. Gales, D. Pye and S. J. Young (1997). Broadcast News Transcription Using HTK.*Proc. of the IEEE-ICASSP*, Vol. 2, 719-722.

IDENTIFICATION DE TRAITS PHONETIQUES ARABES PAR DES SYSTEMES CONNEXIONNISTES MODULAIRES

Sid-Ahmed SELOUANI[1] Jean CAELEN[2]

[1]LCP Institut d'Electronique, USTHB, BP 32 El Alia-Alger-Algérie
Tél: 2 515575 poste 813 e-mail :parole@ist.cerist.dz
[2]CLIPS/IMAG, BP 53, 38041 Grenoble cedex 9 France
Tél: 4 76514627 Fax : 4 76635552 e-mail : jean.caelen@imag.fr

Abstract

This paper deals with an approach which purpose is the improvement of automatic speech recognition systems of Arabic by integrating a mixture of experts. These later are specialized in the identification of specific phonetic features such as the emphasis and the gemination. Experts are typically time delay neural networks using the autoregressive back-propagation algorithm (AR-TDNN). We have assigned to those sub-networks a binary discrimination sub-task in order to attempt an identification of the complex Arabic phonetic features. Serial and parallel architectures of AR-TDNN have been implemented and confronted to a monolithic system.

1. Introduction

La problématique de la reconnaissance automatique de la parole arabe se trouve confrontée en sus des problèmes classiques de la reconnaissance, aux difficultés inhérentes à ses particularités linguistiques. Cependant, avec l'essor et la prédominance des méthodes automatiques, poser aujourd'hui, l'hypothèse de la corrélation des performances d'un système de reconnaissance avec les particularités phonétiques ou phonologiques de la langue utilisée peut paraître inadéquate.

Les performances atteintes par les systèmes actuels [Bou94][Teb95], (Markoviens, neuromimétiques et hybrides) laissent supposer qu'elles sont conditionnées par la seule disponibilité d'un corpus représentatif et convenablement segmenté et étiqueté. Dans le cas particulier de la langue arabe, cette assertion est loin d'être vérifiée. En effet, les produits disponibles sur le marché accusent des insuffisances quant à leur capacité de surmonter les problèmes de forte inflexion de la langue [Ema97]. Les concepteurs des systèmes dédiés à la langue arabe sont unanimes à constater que les consonnes emphatiques, le trait de gémination et l'allongement pertinent des voyelles constituent la source de défaillance principale des systèmes de reconnaissance [Djo89][Ema97].

L'approche que nous présentons ici se base sur des structures de réseaux de neurones à délais et récurrents pour tenter une identification de ces traits phonétiques. Ces réseaux constituent des 'experts' qui viennent affiner l'identification grossière déjà effectuée par d'autres réseaux 'spécialisés' dans l'identification des macro-classes phonétiques. Cette approche 'multi-experts neuronaux' semble prometteuse au regard de l'amélioration du taux de reconnaissance obtenu comparativement aux systèmes opérant une identification frontale.

2. Problématique de l'identification des traits phonétiques arabes

L'originalité de la phonétique arabe se fonde, pour une grande partie sur la pertinence de la durée dans le système vocalique et sur la présence de consonnes emphatiques et du trait de gémination. Ces aspects particuliers jouent un rôle fondamental dans le développement morphologique nominal et verbal.

2.1. L'emphase

Habituellement définie comme un certain type de discours dont les manifestations prosodiques sont liées aux variations d'intonation, l'emphase est dans le cas des langues sémitiques un trait phonétique caractérisant les consonnes. Sur le plan articulatoire, ces consonnes sont réalisées dans la partie antérieure de la cavité buccale, la racine de la langue est reportée en arrière contre la parois pharyngale postérieure et un creusement de la langue est observé.
Il existe dans la langue arabe 4 consonnes emphatiques, 2 plosives : /t̲/, /d̲/ et 2 fricatives : /ð̲/, /s̲/. Dans l'exemple du mot /nasaba/ (a imputé) opposé au mot /nas̲aba/ (a érigé) par la seule emphatisation de /s/, l'ambiguïté est loin de pouvoir être levée par les systèmes actuels même ceux disposant de niveaux linguistiques.

2.2 *La gémination*

Cette particularité compense la pauvreté du système vocalique arabe. Tant au niveau grammatical qu'au niveau sémantique, ce paramètre est fondamental. En effet si l'on observe l'exemple du mot /faʃala/ (a échoué) qui ne s'oppose que par la gémination de la fricative /ʃ/ au mot /faʃ:ala/ (a déjoué), nous mesurons à la fois l'importance et la difficulté d'un système automatique à déceler ce trait.

L'école traditionaliste des grammairiens arabes considère que le trait de gémination est un dédoublement de la consonne (prononcer une consonne d'une manière appuyée induit ce trait sur elle) [Ani70]. Nous avons montré en [Sel96b], confirmant les thèses de Bonnot [Bon79], que l'indice tendu/lâche peut déceler ce trait. Par conséquent, il n'est pas nécessaire d'intégrer lors de l'analyse acoustique le paramètre durée d'une manière explicite.

3. Architectures connexionnistes modulaires et hiérarchisées

Jakobs et Jordan [Jac91] ont introduit une structure hiérarchisée d'experts en vue de résoudre des problèmes de régression non-linéaire. Cette structure s'est révélée dans le cas de la reconnaissance de la parole téléphonique spontanée, plus efficace que les traditionnels réseaux monolithiques [Coo97]. Dans le cas qui nous intéresse à savoir l'identification des macro-classes phonétiques arabes et des traits de gémination et d'emphase, nous proposons des systèmes constitués d'un ensemble de réseaux de neurones simplifiés auxquels nous attribuons des sous-tâches de classification.

3.1 Réseaux de neurones auto-régressifs à délais (AR-TDNN)

La composante temporelle du signal vocal est difficilement modélisable par une approche connexionniste. Cependant, dans le cas particulier de la langue arabe un alignement temporel dans le cas de la durée phonologique (allongement sémantiquement pertinent de la voyelle ou la gémination de la consonne) peut être préjudiciable.

Un système de reconnaissance de l'arabe doit être capable de distinguer entre un allongement dû à une variation de débit par exemple, et celui pertinent, qui est inhérent à une gémination. Les variations du timbre caractérisant les emphatiques et les géminées nous affranchissent de l'utilisation d'un paramètre explicite de durée. Ceci nous a conforté dans l'idée que cette 'perception' doit être préalablement 'apprise' par un système qui 'mémorise' les contextes phonétiques tout en effectuant une intégration temporelle des fenêtres représentant la séquence à identifier. Russel a montré [Rus91] que l'utilisation de l'algorithme auto-régressif de la rétropropagation de l'erreur (AR-backpropagation) permet au réseau d'acquérir cette capacité de mémorisation nécessaire dans le cas de l'identification d'événements temporellement instables. Nous pensons que si nous intégrons dans ce type de réseaux récurrents, une composante de décalage, tel que préconisé dans les réseaux à délais (TDNN : time delay neural networks) [Wai89], nous augmentons la capacité du système à percevoir la durée phonologique même dans un contexte de forte coarticulation. Ceci est illustré par la figure 1.

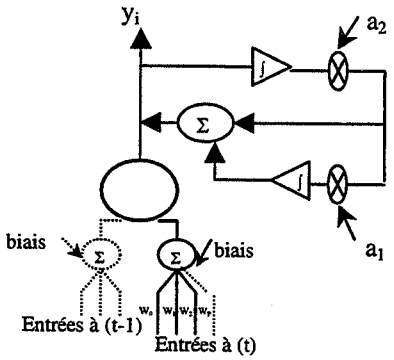


Figure 1. un neurone du réseau AR-TDNN

La réalisation des sous-réseaux intégrant à la fois les propriétés d'une modélisation auto-régressive des sorties des unités cachées et d'un délai sur la couche d'entrée (tel que dans les réseaux TDNN), conduit à une expression de la sortie d'un neurone au niveau de la couche cachée (notée yi) analogue à celle d'un filtre auto-régressif à moyenne ajusté (ARMA).

yi est donnée par l'expression suivante:

$$y_i(t) = f\left(biais_i + \sum_{j=1}^{P} \sum_{m=0}^{L} w_{i,j,m}\ x_j(t-m) \right) + \sum_{n=1}^{M} a_{i,n}\, y_i(t-n)$$

f(x) étant la fonction sigmoïde.
P est le nombre d'unités d'entrées.
L est le nombre d'entrées retardées prises en compte dans le TDNN.
M est l'ordre de prédiction auto-régressive.
Les poids $w_{i,j,m}$ et les biais ainsi que les coefficients $a_{i,n}$ sont les paramètres à optimiser.

Cette optimisation est réalisée par une variante de l'algorithme auto-régressif de rétropropagation de l'erreur donnée en [Rus91], en opérant une intégration temporelle comme pour les réseaux (TDNN) dans la couche d'entrée. Le réseau ainsi constitué sera noté AR-TDNN.

3.2 Structure série

Cette structure est composée d'un agencement sériel de réseaux experts (AR-TDNN). Ces derniers se spécialisant chacun, dans la discrimination d'une classe ou d'un trait phonétique. L'entraînement de chaque expert s'effectue sur tout le corpus d'apprentissage. L'optimisation des

paramètres tels que le nombre de cellules du réseau, la constante d'apprentissage, le nombre et la qualité des entrées ainsi que le nombre d'itérations, est opérée individuellement sur chaque sous-réseau. C'est au moyen d'une validation croisée que se réalise l'ajustement de tous ces paramètres en observant leurs taux de réussite pour différentes valeurs du paramètre à optimiser. Cette tâche est très aisée du fait de la simplicité des sous-réseaux à entraîner.

Deux types de classification de séquences inconnues sont effectuées. L'une grossière, ayant pour objectif la détection des macro-classes : fricatives, plosive, nasale et liquide. La seconde, plus fine, tente de déceler le trait de gémination sur toutes les macro-classes et celui de l'emphase sur les plosives et fricatives. La progression en profondeur (de gauche à droite en figure 2) dans la structure est conditionnée par l'ouverture des portes logiques (et/ou). En effet, selon l'activation des deux sorties d'un réseau expert donné, le processus enclenche l'identification du trait d'emphase ou de gémination si la macro-classe est décelée sinon une activation du réseau adjacent est actionnée. Un échec du système global est comptabilisé si le dernier réseau est atteint sans qu'il y ait discrimination.

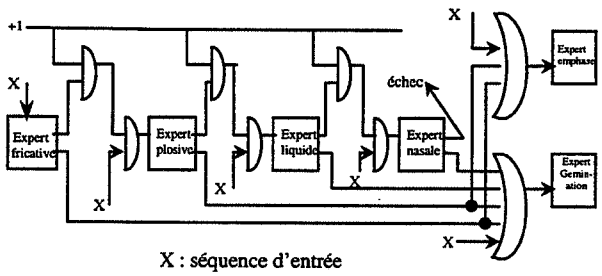


Figure 2 : Structure connexionniste série pour l'identification de macro-classe phonétiques arabes

La structure hiérarchisée (pipe-line) de ce système peut sembler inadéquate dans la mesure où les réseaux situés en profondeur (les plus à droite dans la figure 2) sont pénalisés. Une architecture mettant au même niveau de compétence les réseaux paraît moins contraignante. Cependant, Il faut noter que les experts des niveaux supérieurs (les plus à gauche sur la figure 2) justifient leur position dans le système de test par leur taux de réussite élevé lors de la validation croisée. La pénalisation des niveaux suivants est ainsi minimisée.

3.3 Structure parallèle

Cette configuration est établie sans aucun a priori concernant les performances individuelles relatives de chaque expert. Ces derniers ont le même potentiel d'activation lorsqu'une séquence à identifier se présente à l'entrée. Cette propriété permet une utilisation souple et permet d'éviter le cas d'échec (cf. figure3).

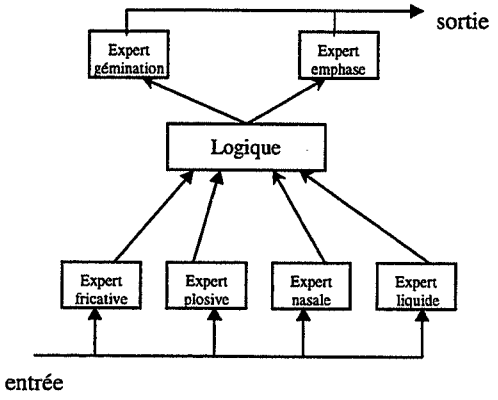


Figure3: Structure connexionniste parallèle pour l'identification de macro-classe phonétiques arabes

Lors de la phase de reconnaissance, il n'est exigé de chaque expert qu'une spécialisation dans le repérage d'une seule (et une seule) classe phonétique. Les sorties activées par les discriminations binaires ainsi opérées sont prises en charge par une logique simple permettant contrairement à la configuration série, de gérer les cas ou plusieurs experts s'activent simultanément. Cette procédure est résumée comme suit :
Considérons $S_{i,1}$ et $S_{i,2}$.
$S_{i,1}$ étant la sortie 1 de l'expert i. Elle se met à l'état +1 si le trait (dont l'identification est dévolue à cet expert i) est détecté. Dans ce cas on dit que l'expert est activé.
$S_{i,2}$ étant 1a sortie 2 de l'expert i. Elle se met à l'état +1 si le trait phonétique ciblé est absent. La classe reconnue est celle dont l'expert est activé. Dans le cas ou le nombre d'experts activés (noté K) est supérieur à 1, alors la classe reconnue est donnée par :

$$\text{Argmax}_j(S_{j,1} - S_{j,2}) \text{ avec } j=1, 2, 3, ...K.$$

Quelque soit l'expert activé, 'l'expert gémination' du niveau supérieur est sollicité. Par contre, l'expert emphase n'est sollicité que si une plosive ou une fricative a été décelée par les niveaux inférieurs.

4. Résultats et commentaires

Le corpus de test a été prononcé par 6 locuteurs (3 hommes et 3 femmes) algériens. Ces mêmes locuteurs ont participé à l'apprentissage et à la validation croisée. Les stimuli sont constitués de 40 occurrences VCV et de 20 phrases (Arabe standard) où les fréquences d'apparition des phonèmes sont respectées.

Le test concerne les 14 fricatives : /f/, /s/, /s̲/, /z/, /h/, /ɳ/, /ʃ/, /θ/, /χ/, /ð/, /ð̲/, /ɣ/, /ɛ/, / /;
les 8 plosives : /t/, /t̲/, /k/, /b/, /d/, /d̲/, /q/, / ʔ/ ;
les 2 liquides :/l/, /r/ ;
les 2 nasales :/m/, /n/ ;
Au total, le test a porté sur 762 fricatives, 592 plosives, 304 nasales et 308 liquides.

Une suite supplémentaire de 219 occurrences VCV dont la consonne est géminée a été testée. Le nombre de consonnes emphatiques (fricatives et plosives) testé est de 184.

Les architectures sérielle et parallèle ont été confrontées à un système monobloc présentant autant de sorties que de classes et de traits à identifier. La segmentation automatique est effectuée par un module fonctionnant sur la base du modèle d'oreille de Caelen [Cae88] et d'une modulation delta d'indices acoustiques [Sel96a]. L'analyse acoustique consiste en une analyse perceptuelle auditive (PLP) [Her90] à laquelle sont ajoutés les paramètres normalisés d'énergie, du taux de passage par zéro et de leur dérivées. A un instant (t), le vecteur acoustique en entrée d'un réseau a une dimension 9.

Au regard des résultats obtenus (cf. tableau 1), il apparaît clairement que les configurations éclatées sérielle et parallèle, avec respectivement 16% et 12% du taux d'erreur moyen, sont plus performantes que la configuration monolithique (28%). Cette amélioration des performances est plus notable dans le cas de la gémination où une diminution de 40% du taux d'erreur a été observée.

Dans le cas des plosives et particulièrement /ʔ/ et /q/ des scores médiocres ont été réalisés par tous les systèmes. Concernant les fricatives, ce sont les fricatives arrières (/h/, /ħ/, /ɣ/, /ɛ/) qui posent le plus de problèmes. Leur brièveté et leur sensibilité au débit (effet de coarticulation) font qu'elles sont le plus souvent fondues dans le contexte vocalique. Les logatomes VCV constituent un matériau défavorable pour l'apprentissage de ce type de fricatives car le pourcentage d'omissions est très élevé. Notons la défaillance totale des systèmes dans l'identification du trait d'emphase pour la consonne /d̲/. l'explication n'est pas dans une difficulté inhérente aux propriétés acoustiques de la consonne, mais plutôt dans la capacité des locuteurs à la prononcer correctement. En effet, dans un contexte VCV, il est très difficile de garder le caractère emphatique de /d̲/ et le plus souvent c'est son opposé par ce trait, /d/ qui est réalisée[1].

Tableau 1. Evaluation par le taux d'erreur des différentes architectures connexionnistes.

Classe / Système	Plosive	Fricative	Nasale	liquide	Emphase	Géminée
Monobloc Simple	23.5 %	22.1 %	25.0 %	28.9 %	29.6 %	38.4 %
AR-TDNN Série	14.0 %	9.8 %	15.7 %	18.3 %	15.9 %	22.7 %
AR-TDNN Parallèle	11.0 %	9.1 %	12.3 %	12.4 %	11.8 %	18.9 %

5. Conclusion

Nous avons présenté les résultats d'identification des macro-classes de l'arabe par deux systèmes à structure hiérarchisée de sous-réseaux de neurones auxquels nous avons alloué des tâches de discrimination binaire (∈ à la macro-classe ou ∉ à la macro-classe). L'approche que nous avons utilisée a pour avantage de faciliter l'apprentissage car la tâche de discrimination binaire ne nécessite pas un grand nombre de cycles (il est de 400) dans la majorité des expériences et chaque sous-réseau est 'initié' indépendamment des autres. La base de données est certes très réduite mais elle est suffisante pour le propos de cet article.

Nos expérimentations ont posé dans le cas de la langue arabe le problème de l'aptitude des systèmes (tout) automatiques de classification (aveugle) à déceler des traits aussi subtils que la gémination et l'emphase. Au regard des résultats obtenus, nous pouvons conclure que les réseaux d'experts connexionnistes dont l'utilisation peut être généralisée à l'identification d'autres traits constituent des systèmes d'appoints simples et performants.

[1] déformation qui est d'ailleurs caractéristique de l'accent régional algérois.

Bibliographie

[Ani70] El-Ani S. (1970), Arabic Phonology, Ed. Mouton The Hague.

[Bon79] Bonnot J.F. (1979), "étude expérimentale de certains aspects de la gémination et de l'emphase en arabe", travaux de l'Institut de .Phonétique de Strasbourg, N°11, pp. 109-118.

[Bou94] Bourlard H., Morgan N. (1994), Connectionist speech recognition: A hybrid approach, Kluwer Publisher.

[Cae88] Caelen J., Tattegrain H. (1988), « Le Décodeur Acoustico-Phonétique dans le projet DIRA », XIIèmes JEP, Nancy , pp. 115-121.

[Coo97] Cook G.D, Waterhouse S.R. (1997) « Ensemble methods for connectionist acoustic modeling » Eurospeech97, Rhodes, Greece, pp. 1959-1962.

[Djo89] Djoudi M., Fohr D., Haton J.P. (1989), « Phonetic study for automatic recognition of Arabic », European Conference on speech and technology, pp. 268-271.

[Ema97] Emam O. (1997) « Speech recognition of Arabic », Technical notice on www-page of IBM Cairo scientific center.

[Her90] Hermansky H. (1990), "perceptual linear predictive (PLP) analysis of speech", JASA journal, 87 (4), pp. 1738-1752.

[Jac91] Jacobs R.A, Jordan M.I, Nowlan S.J., Hinton G.E. (1991), "Adaptative mixtures of local experts", Neural computation, 3(1) pp. 79-87.

[Rus91] Russel R. L., Bartley C.(1991), « The autoregressive backpropagation algorithm », International Joint Conference on neural networks Vol II pp. 369-377.

[Sel96a] Selouani S.A, Caelen J.(1996), « Experiments on arabic phone recognition using automatically derived indicative features », in proceedings of IVth ISSPA, Gold coast, Australia.

[Sel96b] Selouani S.A, Caelen J.(1996), « validation de traits phonétiques par un système de reconnaissance de l'arabe standard », XXI JEP, Avignon, pp. 347-350.

[Teb95] Tebelskis J. (1995), Speech recognition using neural networks, PHD thesis, Carnegie Mellon University.

[Wai89] Waibel A., hanazawa T., Hinton G., Shikano K. (1989), "Phoneme recognition using time-delay neural networks", IEEE trans. on ASSP, Vol. 37(3) pp.328-339.