ACTES

XX èmes JOURNEES D'ETUDE SUR LA PAROLE

TREGASTEL, du 1er au 3 JUIN 1994

Sous le patronage du GFCP,
groupe spécialisé de la SFA/ESCA,
avec l'aide de l'APAST

et de France Telecom
CNET

De Mori

France Télécom
CNET
Groupement TSS

XX èmes
JOURNEES D'ETUDE
SUR LA PAROLE

XXèmes JOURNEES D'ETUDE SUR LA PAROLE

Organisées sous le patronage du GFCP,
groupe spécialisé de la SFA/ESCA, avec
l'aide de l'APAST et de

France Telecom
CNET

Dédié à Max WAJSKOP

Groupement TSS
France Télécom CNET/LAA
Route de Trégastel
22300 Lannion

Tél : +33 96 05 30 88
Fax : + 33 96 05 14 12

TABLE DES MATIERES

Paroles d'anniversaire..... I
Louis-Jean BOE

Mot du Comité d'Organisation...... II

Allocution d'ouverture

P. LORAND (Chef du SCL au CNET Lannion)..... III

COMMUNICATIONS INVITEES..... 1

→ Sciences de la parole : genèse d'une communauté scientifique en France..... 3
M. GROSSETTI
→ Apprentissage automatique pour l'interprétation sémantique..... 11
R. DE MORI

SESSION ORALE O1 : HISTOIRE DES SCIENCES DE LA PAROLE..... 21
Président de Séance : M. GROSSETTI

- → La recherche publique française et les premiers ordinateurs : un échec lourd de conséquences (1939-1959)..... 23
  (P.E. MOUNIER-KUHN)
- → Télécommunications et traitement de la parole : quelques éléments historiques..... 31
  (M. CARTIER - C. GAGNOULET)
- → Le premier calcul des séries de Fourier..... 37
  (J.B. ROBERT)
- → La modélisation articulatoire : un demi siècle d'évolution entre fonctionnel, physique et biomécanique..... 41
  (L.J. BOE - S. MAEDA - P. PERRIER)

SESSION ORALE O2 : PROSODIE..... 55
Président de Séance : C. SORIN

- → Patrons prosodiques et intentions des locuteurs : production et perception de formes expressives chez l'adulte et l'enfant..... 57
  (C. GERARD - J. CLEMENT)
- → Etude de l'influence des variations prosodiques dans les processus de traitement de la parole..... 63
  (D. DAHAN)
- → Synthèse de parole de haute qualité à partir d'un texte : une comparaison de quatre algorithmes candidats..... 69
  (T. DUTOIT - H. LEICH)
- → Recherche d'une stylisation extrême des contours de F0 en vue de leur apprentissage automatique..... 75
  (S. de TOURNEMIRE)
- → La perception de l'accent en Français du Québec et en français de France..... 81
  (D. DESHAIES - I. GUAITELLA - C. PARADIS)
- → Caractérisation des variations individuelles du contour de fréquence du fondamental observées dans des phases lues en anglais..... 87
  (J. VAISSIERE)

SESSION AFFICHEE A1 : PRODUCTION ET SYNTHESE 93

→ Vers une phonémisation automatique des sigles 95
(P. BOULA de MAREUIL)
→ Formalisation de connaissances prosodiques à l'aide d'un système d'apprentissage symbolique et numérique 101
(J.J. SCHNEIDER - P. ROBINEAU - H. MELONI)
→ Contrôleur gestuel des cordes vocales d'un robot parlant 107
(C. VESCOVI - E. CASTELLI)
→ Formulation variationnelle du lien acoustico-articulatoire 113
(P. JOSPA)
→ Adaptation d'un modèle articulatoire à un locuteur, dans le but de contraindre l'inversion articulatori-acoustique 119
(D. BEAUTEMPS - B. GABIOUD)
→ Le compilateur de règles compost et son application pour l'analyse syntactico-prosodique dans un système de synthèse Psola en allemand 125
(M. ALISSALI - S. FRENKENBERGER - M. KOMMENDA - B. SCHNABEL)
→ Comparaison entre la variabilité des formants et la variabilité d'estimations des cibles formantiques 131
(M. PITERMANN - J. CAELEN)
→ Protection de segments sub-phonétiques en synthèse PSOLA 137
(P.Y. LE MEUR)
→ Un modèle 3D de lèvres parlantes
(A. ADJOUDANI - T. GUIARD-MARIGNY - B. LE GOFF - C. BENOIT) 143
→ Analyse synthèse de visages parlants 147
(O. ANGOLA - B. LE GOFF - T. GUIARD-MARIGNY - A. ADJOUDANI - C. BENOIT - M. COHEN)
→ L'écart entre F1 et F2 comme mesure de l'effet des consonnes linguales sur les cibles acoustiques des voyelles de l'arabe standard moderne 215*
(I. ZNAGUI - A. BOUCHNAFI)

SESSION AFFICHEE A2 : SEGMENTATION ET RECONNAISSANCE 151

→ Utilisation d'un réseau neuro-flou pour la distinction du trait voisé/non-voisé 153
(R. SOKOL - P.Y. GLORENNEC - G. MERCIER - K. WOLINSKI)
→ Comparaison des multimodèles et des densités multigaussiennes pour la reconnaissance de la parole par modèle de Markov 159
(D. JOUVET - M. DAUTREMONT - A. GOSSART)
→ Une méthode temporelle de discrimination bruit/parole en milieu bruité 165
(J. B. PUEL - R. ANDRE-OBRECHT)
→ Deux systèmes d'étiquetage phonétique basés sur une pré-segmentation du signal de parole 169
(A. FARHAT - N. PARLANGEAU - R. ANDRE-OBRECHT)
→ Comparaison de deux méthodes non paramétriques pour la détection des occlusives sourdes 175
(F. MALBOS - R. ANDRE-OBRECHT - M. BAUDRY)

* Communication insérée par erreur dans les actes en session affichée A2.

→ Choix et adaptation d'un phonétiseur pour la reconnaissance automatique de la parole....181
(K. BARTKOVA - D. LARREUR - I. METAYER)
→ Un algorithme pour l'estimation des instants de fermeture de la glotte à l'aide de la détection séquentielle des ruptures du signal de parole....187
(C. MURGIA - G. FENG)
→ Evaluation de la reconnaissance de traits phonétiques en utilisant l'information mutuelle....193
(C. WINDHEUSER - S. OKAWA - F. BIMBOT - K. SHIRAI)
→ Modèles stochastiques d'ordre 1 et 2....199
(J. F. MARI - D. FOHR - J.P. HATON)
→ Efficacité des méthodes diagnostiques de l'insuffisance vélaire sur segments vocaliques....203
(E. PERRIN - C. BERGER-VACHON - F. PLANTE - I. KAUFFMANN)
→ Un environnement logiciel pour un atelier phonétique....209
(Y. LAPRIE - L. MERCIER)

SESSION ORALE O3 : TRAITEMENT DU SIGNAL....221
Président de Séance : M. BAUDRY

→ Adaptation aux conditions des lignes téléphoniques d'un système de reconnaissance automatique de la parole....223
(C. MOKBEL - P. PACHES-LEAL - J. MONNE - D. JOUVET)
→ Approche coopérative de l'extraction de la fréquence fondamentale....229
(A. SOQUET)
→ Extraction des formants en synchronie avec l'excitation à l'aide d'un modèle auto-régressif composé....235
(Z. AZAMI - J. SCHOENTGEN)
→ Codage de parole CELP à excitation vectorielle codée par treillis
(A. POPESCU - C. LAMBLIN - N. MOREAU)....241
→ Localisation de sources virtuelles par synthèse binaurale : Simulation d'une écoute stéréophonique....247
(M. MARIN - A. GILLOIRE - J. F. LACOUME - O. WARUSFEL - J.M. JOT)

SESSION AFFICHEE A3 : BASES DE DONNEES ET DE CONNAISSANCES....253

→ Investigation de protocoles pour la reconnaissance des locuteurs fondée sur le système Orphée....255
(J. L. LE FLOCH - C. MONTACIE - M.J. CARATY)
→ Avantages d'une approche analytique orientée connaissances en reconnaissance du locuteur....259
(J.F. BONASTRE - H. MELONI)
→ Recherches comparatives pour caractériser des locuteurs, en particulier des jumeaux....265
(W. ENDRES - M. DEPPE - M. FISCHER)
→ Etude des variabilités inter et intra-locuteurs provoquées par l'émotion et l'imitation pour les systèmes de vérification du locuteur....269
(M. HOMAYOUNPOUR - G. CHOLLET - J. VAISSIERE)
→ Approches prédictives neuronales pour l'identification....275
(T. ARTIERES - P. GALLINARI)
→ Le développement et l'enregistrement de ICY, un corpus pour l'étude de la variabilité Intra-locuteur et la Caractérisation de stYles parlés....281
(V. PEAN - M. ESKENAZI)

→ BDBRUIT Une base de données "Parole" de locuteurs soumis à du bruit..... 287
(J. ZEILIGER - J.F. SERIGNAT - D. AUTESSERRE - C. MEUNIER)
→ Test et mise au point des paquets de règles de production pour la communication parlée..... 291
(P. MUNTEANU - J.F. SERIGNAT)
→ Lemmatisation du dictionnaire phonétique du français parlé "de A à ZUT"..... 297
(D. DUJARDIN - R. BELRHALI - L.J. BOE - J. COURTIN)
→ Sauvegarde lexicale et phonétique de dialectes, origine des toponymes..... 303
(A. FRANCES - R. BELRHALI)
→ Transcription multiniveau d'un corpus de parole..... 309
(A. MAILLAND - M. DE CALMES - G. PERENNOU)
→ Catégorisation phonétique et subphonétique dans un lexique du français Etude préliminaire 315
(H. AIT-MESSAOUD - V. AUBERGE - L.J. BOE)
→ Utilisation conjointe de connaissances et de techniques neuromimétiques pour la reconnaissance du locuteur..... 321
(R. QUELAVOINE - P. NOCERA - J.F. BONASTRE)
→ Les phénomènes de variation de geste : Etude des variations articulatoires de groupes consonantiques sous l'effet du débit, en Français..... 325
(B. VAXELAIRE)

SESSION AFFICHEE A4 : PHONETIQUE ET PROSODIE..... 331

→ The phonetic typology of germanic long and short vowels..... 333
(R.K. POTAPOVA)
→ La segmentation de la parole chez des adultes lettrés et illettrés arabophones : approche psycholinguistique..... 335
(M. FARID)
→ Caractères spécifiques des voyelles nasales du Français..... 339
(J.S. LIENARD - M.G. DI BENEDETTO)
→ Matrices phonétiques et matrices phonologiques arabes..... 345
(B. BOUDRAA - S.A. SELOUANI - B. GUERIN - M. BOUDRAA)
→ Oralité, contexte et cognition sur le terrain créole réunionnais..... 351
(G. STAUDACHER-VALLIAMEE)
→ Le rôle des similarités acoustico-phonétiques dans l'accès au lexique chez le bilingue arabe/anglais : preuves à partir du Gating..... 355
(S. BOUDELAA)
→ Vers la perte de l'opposition du lieu d'articulation des voyelles moyennes arrondies dans la parole des jeunes gens d'Ile-de-France..... 361
(I. MALDEREZ)
→ Faits systématiques de durée en Français Québécois, rythme et accentuation..... 367
(M. OUELLET)
→ La phonétisation du Français : logique et linguistique dans une grammaire-lexiques..... 373
(R. BELRHALI - V. AUBERGE - L.J. BOE)
→ Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en Français..... 379
(E. DELAIS)
→ Intérêt du test de rimes dans l'évaluation de l'intelligibilité en présence de bruit masquant..... 385
(J.M. BENE - C. DUBREUIL - J.C. DUCLOS - J.C. NORMAND - M. VERNET - G. PACHIAUDI)

→ Interaction entre l'activité gestuelle et l'activité vocale dans la communication : éléments théoriques et méthodologiques 389
(I. GUAITELLA)
→ L'opposition de durée vocalique en Arabe : essai de typologie 395
(J. MOUNIR)
→ Tests de hauteur et perception de l'accent en anglais par des francophones : une première approche 401
(A. DIANA)

SESSION ORALE O4 : PRODUCTION 407
Président de Séance : H. MELONI

→ Pharyngales et uvulaires arabes sont des approximantes : caractérisation acoustique 409
(M. YEOU - S. MAEDA)
→ Quelles stratégies de compensation pour une perturbation labiale ? 415
(C. SAVARIAUX - P. PERRIER - J.P. ORLIAGUET)
→ La coarticulation mandibulaire comme principe d'organisation syllabique (Et si la mandibule syllabisait ALIS ?) 421
(N. RHARDISSE - C. ABRY)
→ Estimation de trajectoires articulatoires à l'aide d'un réseau de neurones 427
(M. GEORGE - P. JOSPA - A. SOQUET)
→ Effet de la connaissance de la voyelle sur la perception du bruit d'explosion des occlusives 433
(L. DJEZZAR - A. BONNEAU - Y. LAPRIE)
→ Modélisation des mouvements articulatoires par la méthode de la MLPC et ses effets sur les caractéristiques acoustiques 439
(S. BOUABANA - S. MAEDA)

SESSION ORALE O5 : PHONETIQUE 445
Président de Séance : L.J. BOE

→ Les facteurs responsables de l'assimilation : analyse de trois types de groupes de consonnes et de leur miroir 447
(C. MEUNIER)
→ Corrélats acoustiques et perception de la diphontgaison des voyelles françaises /e/ et /o/ par les locuteurs américains 453
(B. LAURET)
→ Les consonnes "fortes" du berbère : indices perceptuels et corrélats phonétiques 459
(N. LOUALI - G. PUECH)

SESSION ORALE O6 : RECONNAISSANCE 465
Président de Séance : R. DE MORI

→ Modèles de séquences à horizon variable : multigrams
(F. BIMBOT - R. PIERACCINI - E. LEVIN - B. ATAL) 467
→ Interaction entre le décodage acoustico-phonétique et l'analyse lexicale grâce aux techniques de raisonnement hypothétique : le système DAPHNE 473
(S. COSTE-MARQUIS)
→ Détection de mots clés dans un discours continu 479
(F. BECHET - S. IGOUNET - H. MELONI - P. GILLES)

→ Inversion du modèle DRM pour la reconnaissance de la parole : Application à l'identification de diphones vocaliques.....485
(L. CANDILLE - H. MELONI - T. SPRIET - R. CARRE)
→ Topologie des locuteurs et présegmentation pour le décodage acoustico-phonétique.....491
(C. BARRAS - M.J. CARATY - C. MONTACIE)
→ Reconnaissance de mots enchaînés par une méthode hybride : réseaux markoviens et base de règles.....497
(O. DELEMAR)

SESSION ORALE O7 : PROSODIE ET PSYCHO-PERCEPTION 503
Président de Séance : W. SERNICLAES

→ Accent, durées syllabiques et segmentation lexicale.....505
(M.H. BANEL - N. BACRI)
→ Leçon de déclinaison de Marguerite DURAS et de Marguerite YOURCENAR.....511
(Z. FAGYAL)
→ Segmentation lexicale en Français parlé - La liaison enchaînée.....517
(S. WAUQUIER-GRAVELINES)
→ La chaîne phonétique et le patron accentuel lexical sont-ils représentés séparément ? Une étude d'illusions perceptives.....523
(V. PASDELOUD - R. KOLINSKY - J. MORAIS)
→ L'Effet du débit de parole sur le rôle du point d'unicité.....529
(M. RADEAU - P. MOUSTY - V. PASDELOUP)
→ Segmentation et reconnaissance des mots parlés en contexte : une approche computationnelle.....535
(A. CONTENT - P. STERNON)

Liste alphabétique des auteurs.....541

«La jeunesse n'aime pas les objets parfaits,
ils lui laissent trop peu à faire [...] et l'ennuient...»

Que l'on se rassure les jeunes chercheurs,
la parole est un objet beaucoup trop humain pour être parfait !

PAROLES D'ANNIVERSAIRE...

«J'ai vingt ans et je ne laisserai à personne dire que c'est le plus bel âge de la vie»... Permettez-moi, permettons-nous, linguistes, informaticiens, phonéticiens, dialectologues, électroniciens, psychologues, phonologues, acousticiens, traiteurs de signaux, spécialistes du dialogue, d'Intelligence Artificielle, de Sciences Cognitives, et pour tout dire de parole... de ne pas être tout à fait de l'avis de l'écrivain. Sans vouloir filer outre mesure la métaphore entre la juvénilité éclatante de vingt printemps et la toute jeune maturité du Groupe Francophone de la Communication Parlée, c'est avec une certaine fierté que l'on peut rétrospectivement apprécier ces vingt JEP écoulées.

Commençons par une interrogation géographique. Est-ce un hasard si les JEP sont organisées pour la troisième fois par le CNET, haut-lieu des télécommunications ? Nous ne le croyons pas, et nous renvoyons à la conférence invitée de Michel Grossetti pour nous remémorer, s'il était besoin que :
(1) le téléphone a été la première technique initiatrice pour la parole, attirant dans son champ des ingénieurs motivés par les transmissions ; fascinés, certains d'entre eux ont recentré leur recherche sur l'objet à transmettre (ce n'est pas Gunnar Fant qui nous dirait le contraire) ;
(2) le CNET a joué en France, dès le début des années 60, un rôle décisif en motivant des équipes, sollicitant des projets et distribuant (avec IBM La Gaude) des échantillons de parole numérisée, à une époque où les premiers foyers de traitement numérique de la parole ne disposaient pas d'entrée/sortie de type vocodeur ;
(3) le CNET a marqué l'organisation de la communauté Parole, de la fin des années 70 jusqu'au milieu des années 80, en favorisant les recherches fondamentales et finalisées, et en réalisant toute la mise en œuvre de la base de données des sons du français. Même si la modestie de Raymond Descout, l'ancien responsable du département RCP et l'ancien président du G(F)CP, doit en souffrir, nous lui devons beaucoup.

Avançons maintenant une hypothèse : les Sciences de la Parole se sont développées initialement aux USA au début des années 50, au MIT et aux laboratoires Haskins. Le manifeste annonciateur en est vraisemblablement les *«Preliminaries to Speech Analysis»*. Les auteurs R. Jakobson, G. Fant et M. Halle - un linguiste, un spécialiste de l'acoustique et des télécommunications et un phonologue - situent le cadre transdisciplinaire permettant de passer des traits linguistiques au signal. Avec *«Some Experiments on the Perception of Synthetic Speech Sounds»*, sont mis en évidence, par une équipe d'électroniciens,

phonéticiens, psychologues, (F.S. Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst et L.J. Gerstman) les indices acoustique responsable de la perception des sons. C'est l'objet de l'étude, *speech* (la parole), qui va fixer l'appellation de la finalité d'une nouvelle discipline, qui n'est ni tout à fait de l'acoustique... En France c'est à la fin des années 60 que sont réunies les conditions de son émergence ; c'est le Groupe Communication Parlée qui est le creuset dans lequel vont se fondre les heureux ingrédients du mélange. Au terme d'une présidence de quatre ans (de 19756 à 1978), Max Wajskop fait acquérir au groupe la stature d'une société scientifique dotée d'une structure, d'un bureau et d'objectifs. Il dote la communauté parole d'une revue européenne, *Speech Communication* ; saluons ici les choix visionnaires de notre ami disparu en 1993 qui était encore parmi nous, comme organisateur des dernières JEP de Bruxelles.

Les JEP constituent des moments forts et des repères datés, dont les actes retracent les évolutions de notre discipline. La richesse des thèmes abordés, le nombre des communications retenues cette année par le comité scientifique et leur qualité témoignent du chemin parcouru. Vingt éditions des JEP, presque 25 ans d'existence pour le GFCP, est venu le temps de nous pencher sur notre histoire, sur l'histoire des Sciences de la Parole, en approfondissant la réflexion initiée en 1989 aux JEP de Nancy. Notre Groupe a eu beaucoup de chance en pouvant s'assurer de la collaboration d'un sociologue, spécialiste de la répartition territoriale des Sciences de l'Ingénieur et d'un historien de l'Informatique. Vous trouverez dans ces Actes, au thème consacré à l'Histoire de la Parole, les contributions de Michel Grossetti et de Pierre E. Mounier-Kuhn prolongées par les travaux de la Commission Parole et Informatique du GFCP et du Groupe de Travail (GT0) du PRC Informatique (responsable Jean Caelen).

Nous avons commencé en citant Paul Nizan, terminons en répondant, pour le GFCP à une des interrogations du poète :

«Dis, qu'as-tu fait, toi que voilà,
De ta jeunesse ?»

De la parole, pour sûr !

Louis-Jean Boë
Ancien Président et
membre du bureau de GFCP...
depuis 20 ans

BIENVENUE AUX XXèmes JEP

Pour vous accueillir à Trégastel et célébrer cet anniversaire des JEP, comment ne pas citer le charmant (et indispensable) texte de David Lodge :

" Le congrès de l'ère moderne ressemble au pélerinage des chrétiens du Moyen Age dans la mesure où il permet à ses participants de se livrer à tous les plaisirs et à tous les divertissements inhérents au voyage sans qu'ils aient à se départir de cet air austère qui sied à toute recherche pour se mieux connaître. Reste bien sûr que certains exercices pénitentiels doivent être accomplis : faire une communication, peut-être, et naturellement écouter celles des autres. Mais, grâce à cet alibi, vous visitez des lieux nouveaux et intéressants, vous rencontrez des gens nouveaux et intéressants, et créez avec eux des relations nouvelles et intéressantes ; vous échangez ragots et confidences (car vos histoires éculées sont neuves pour eux, et vice versa) ; vous mangez, buvez et vous divertissez en leur compagnie chaque soir ; et pourtant, quant tout est fini et que vous rentrez chez vous, votre sérieux intellectuel ne s'en trouve qu'accru. Les congressistes d'aujourd'hui ont un avantage supplémentaire sur les pélerins d'autrefois : leurs frais sont généralement payés, ou en tout cas couverts en partie grâce à des subventions versées par l'institution à laquelle ils appartiennent, qu'il s'agisse d'un département ministériel, d'une firme commerciale, ou plus communément peut-être, d'une université."

Bienvenue à Trégastel, donc.

Heureuses JEP à tous, auteurs et participants.

Merci à tous ceux, individus, institutions qui ont permis leur organisation.

Le Comité d'Organisation

XXèmes JOURNEES D'ETUDE SUR LA PAROLE
Trégastel (Côtes-d'Armor) - 1er au 3 juin 1994

ALLOCUTION DE BIENVENUE

Pierre LORAND
Chef du Service Commun Lannionnais du CNET
Ancien Président du GALF

Mesdames, Messieurs,

Le Président du Comité d'organisation, André SOUBIGOU, m'a demandé de prononcer l'allocution inaugurale des XXèmes Journées d'Etude sur la Parole ; c'est un honneur et un grand plaisir pour moi de souhaiter la bienvenue en Bretagne, dans ce site magnifique de Trégastel, aux nombreux participants, dont certains viennent de très loin. Souhaitons que le temps soit assez clément pour que le plaisir des joies touristiques s'associe aux satisfactions scientifiques attendues d'une telle rencontre.

C'est un honneur de vous accueillir au nom du Comité d'organisation, Pierre DUPONT, Michel CARTIER, Laurent MICLET, et de son Président ; de représenter ici la Direction du CNET, qui tient à marquer l'importance que le centre de recherches de France Télécom attache à cette manifestation et souligner l'utilité de ces échanges réguliers de la communauté internationale des chercheurs francophones spécialistes de la parole.

C'est aussi un grand plaisir personnel d'avoir l'occasion de saluer des anciens collègues que j'ai connus dans les années 1970-1980, au moment où j'avais la responsabilité d'animer les équipes du CNET travaillant dans le domaine de l'acoustique et du traitement du signal de parole, et où je participais intensément à la vie du GALF - devenu depuis la SFA - en tant que Secrétaire Général puis comme Président. Revoyant des visages connus, je regrette pourtant, comme vous tous, l'absence d'un chercheur éminent qui a marqué de son empreinte la vie du Groupe Communication parlée, Max WAJSKOP, disparu prématurément l'an dernier, et je salue ici sa mémoire.

Pouvoir fêter la XXème édition des Journées d'Etude sur la Parole prouve à quel point est vivante la communauté des spécialistes qui s'intéressent au moyen de communication le plus naturel de l'homme. Je n'insisterai sur les aspects historiques, car certains exposés de ce matin vous rappelleront la vie scientifique des trente dernières années dans ce domaine. Je veux toutefois souligner la belle régularité, annuelle, puis bisannuelle, de ces rencontres inaugurées en 1970 à Grenoble, qui se sont tenues depuis dans de nombreux sites français, mais aussi hors de nos frontières, à Bruxelles, Montréal et Hammamet.

Lannion avait accueilli les 3èmes JEP en 1972, et les 9èmes en 1978...il est bien naturel que la Bretagne accueille à nouveau cette manifestation ; la tenue de la 20ème édition à Trégastel est un symbole pour les équipes du CNET-Lannion, symbole de la maturité de leurs recherches et des espoirs qu'apporte "la force de l'âge".

L'effort de recherche du Centre National d'Etudes des Télécommunications, depuis les premières études sur les vocodeurs à canaux en 1960, ne s'est, en effet, jamais relâché dans le domaine du traitement du signal de la parole. C'est normal, penserez-vous, puisque la parole constitue aujourd'hui et pour longtemps encore la plus grande partie des informations transmises par un réseau de télécommunications. Mais le changement de statut de l'ancienne Direction Générale des Télécommunications de l'Administration des PTT, qui est devenue France Télécom, opérateur soumis à la concurrence nationale et internationale, peut faire évoluer la nature des recherches du CNET et suscite parfois des craintes pour l'avenir. Il est vrai que les programmes de recherche du CNET sont progressivement orientés vers des études plus directement appliquées aux télécommunications, en particulier pour la création de nouveaux services offerts aux clients. Mais, loin de restreindre le domaine de recherche, les études de nouveaux services ouvrent de larges perspectives à l'équipe du CNET à Lannion : la reconnaissance de la langue, l'authentification vocale, la reconnaissance de la parole monolocuteur avec apprentissage, à hautes performances, sont des axes qui doivent être explorés pour leurs applications dans de nouveaux services ; parallèlement, les études relatives au codage et à la prise de son (antennes acoustiques, "débruitage"...) doivent être poursuivies, compte tenu de leur importance pour les communications avec les mobiles, en plein développement.

Ne pouvant tout faire elles-mêmes, les équipes du CNET doivent aussi s'appuyer sur une recherche à caractère plus "amont", développée dans le milieu scientifique. Cette collaboration s'effectue grâce à des conventions de partenariat avec le CNRS, les universités, les laboratoires des écoles..., et dans ce but, le Service de l'Action

Scientifique du CNET a lancé des consultations thématiques en direction de ces partenaires.

Je crois que l'avenir des chercheurs de la Communication parlée peut être envisagé avec optimisme. Les applications pratiques de leurs études, qui apparaissent aujourd'hui dans la vie quotidienne des citoyens, qu'il s'agisse de codage, de synthèse ou de reconnaissance de parole, justifie les efforts considérables consentis depuis trente ans par la communauté scientifique et technique ; ces premières applications appellent des perfectionnements ultérieurs, suscitent de nouvelles pistes de recherche et permettent d'imaginer, en conjugaison avec les progrès fantastiques de la technologie, qu'un jour la machine approchera les performances de l'homme, pour la production, la perception et le traitement du signal de parole.

Je souhaite un plein succès aux XXèmes Journées d'Etude sur la Parole et, à tous, un agréable séjour sur la Côte de Granit Rose. Je suis certain que vous reviendrez dans ce beau pays pour des futures JEP...Pourquoi pas les 30èmes ?

Merci de votre attention.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

COMMUNICATIONS INVITEES

Sciences de la parole : genèse d'une communauté scientifique en France[1]

Michel Grossetti
Centre de Recherches Sociologiques
et Centres Interdisciplinaire d'Etudes Urbaines
Université de Toulouse le Mirail

En 1970 se tenaient à Grenoble, dans le cadre du Groupement des Acousticiens de Langue Française (GALF), les premières Journées d'Études sur la Parole (JEP). Cette première manifestation allait par la suite se pérenniser puisque nous assistons en 1994 à la vingtième édition de cette rencontre. La création des JEP en 1970 est un événement important dont la signification doit être interrogée. S'agit-il d'une nouvelle manifestation parmi d'autres au sein d'un domaine scientifique déjà institué ou a-t-on affaire au contraire à la constitution d'une nouvelle communauté de recherche ? Bien sûr, les JEP ne sont pas la première occasion pour les chercheurs de se rencontrer autour du thème de la parole. En 1967 avait été organisé, déjà dans la cadre du GALF et à Grenoble, un colloque sur les « Structures acoustiques de la parole ». Ce colloque est certainement la première manifestation associant en France des spécialistes de la phonétique, de l'électronique, de l'informatique et des pathologies de l'audition ou du langage. Il n'a toutefois pas eu de véritable suite, si ce n'est précisément les JEP, dont il constitue une sorte d'avant première. On ne peut pas considérer non plus qu'il n'y a qu'un changement d'intitulé entre les deux manifestations puisque le comité d'organisation des premières JEP ne comprend plus aucun de ceux qui avaient conçu le colloque de 1967, même si les institutions impliquées sont sensiblement les mêmes. Les liens qu'entretiennent ces deux événements seront détaillés dans la suite de cette communication, mais l'on peut admettre provisoirement que les JEP constituent bien la première forme régulière d'institutionnalisation d'une communauté scientifique alors en émergence, celles des sciences de la parole, ce qui forme le point de départ de la réflexion proposée ici.

Pour qu'une manifestation de cette ordre puisse se créer et perdurer, il est nécessaire qu'elle suscite l'intérêt de suffisamment de chercheurs et d'équipes autour d'un même thème. C'était le cas en 1970, ce qui signifie qu'existait alors en France, au moins de façon latente, les éléments d'une communauté scientifique. Comment le thème de la parole s'est-il développé dans les diverses disciplines ? Comment les chercheurs travaillant sur ce thème ont-ils été amenés à entrer en contact les uns avec les autres ? Pourquoi est-ce au tournant des années soixante-dix que cette communauté prend forme ? En définitive, qu'est ce qui rend possible la tenue des JEP de 1970 ? Telles sont les questions traitées dans cette communication, du point de vue d'un sociologue, ce qui signifie que l'on s'intéressera plus aux institutions, aux hommes et à leurs interactions, qu'aux contenus scientifiques eux-mêmes. C'est donc une analyse « externaliste » de la genèse de la communauté des sciences de la parole en France qui est proposée ici.

Pour cela, il est nécessaire de revenir rapidement sur la situation en France dans la période 1945 - 1970 des diverses disciplines qui sont concernées par les premières JEP (Phonétique, Électronique, Informatique), ainsi que des principaux organismes de recherche technologique amenés à jouer un rôle dans le domaine de la parole. Ensuite est analysé le processus par lequel des chercheurs ou des équipes sont amenés à converger vers l'objet parole. Ce processus fait intervenir des logiques variables : émergence de l'objet dans une discipline générique, passage d'une définition de l'identité scientifique par les méthodes (informatique par exemple) à une définition par l'objet lui-même, etc. La réflexion se focalise sur une dimension particulière de l'émergence de la communauté des sciences de la parole, qui est le jeu des configurations institutionnelles et scientifiques locales dans la mise en relation des différents acteurs.

1 - Prémices

Avant que ne débute en France le processus de structuration qui amènera la création des JEP, ce que nous appelons actuellement la parole avait fait l'objet de différentes approches et connu des évolutions importantes dans les techniques d'enregistrement, de reproduction et de transmission.

1.1 - Les tribulations d'un objet interdisciplinaire

L.J. Boë et J.S. Liénard (1988, 1989) et L.J. Boë (1990) ont décrit comment la parole a successivement intéressé à partir du XVIIIe siècle

la physiologie attachée à décrire les mécanismes de la parole pour prévenir ou soigner les pathologies qui peuvent s'y rattacher, la physique qui cherche à enregistrer et reproduire le signal, et la phonétique qui constitue la parole en objet pour les sciences de l'Homme. Avec nos catégories actuelles, on pourrait dire que se trouvent donc impliquées à différents moments dans des travaux sur la parole les sciences de la matière, les sciences de la vie et les sciences de l'homme. Ce caractère d'objet pouvant être abordé depuis des rivages diciplinaires très éloignés au départ est probablement l'une des particularités fortes de la parole. Il est significatif que ce soit au sein de la phonétique, discipline universitaire qui a placé depuis le début la parole au centre de ses préoccupations, que l'aspect interdisciplinaire de l'objet soit apparu le plus rapidement, comme le montre l'organisation en 1932 du premier Congrès International de Sciences Phonétiques qui opère une rupture avec la phonétique en affirmant la pluralité des approches possibles de la parole. Cette pluralité s'exprime dans les thèmes du congrès qui vont de la physiologie aux expérimentations acoustiques et comprennent même une session de démonstration d'appareils de mesure ou d'enregistrement. Objet complexe de disciplines différentes, la parole est aussi un objet pour les ingénieurs et la recherche industrielle, à travers l'invention du téléphone ou du phonographe bien sûr, mais aussi au cours des années trente et quarante, avec les avancées de l'instrumentation électro-mécanique (*Vocoder* en 1936, *Voder* en 1939, *Sonagraph* en 1946, Cf. Boë et Liénard, 1988). Il semble qu'à partir du moment où l'on dispose d'une insrumentation susceptible d'enregistrer et de reproduire la voix sous une forme graphique, des chercheurs du domaine des sciences de l'ingénieur[2] (électroniciens en particulier) commencent à travailler régulièrement sur la parole aux États-Unis. Parallèlement à l'évolution des matériels, se met en place et se diffuse une conceptualisation du phénomène parole distinguant une fréquence fondamentale et des formants qui, en modulant cette fréquence, produisent ce qui va véhiculer le sens. Cette émergence progressive d'un corpus de connaissances et de méthodes d'étude trouve sa première formulation synthétique dans l'ouvrage, « Preliminaries to speech analysis » (1952), dont il est significatif qu'il soit signé par un linguiste, R. Jakobson, un électronicien, G. Fant, et un phonologue, M. Halle. Les auteurs proposent dans cet ouvrage une théorie universelle des rapports entre la structure linguistique et le son. Il est tout aussi significatif que celui des trois auteurs qui prolongera cette réflexion (G. Fant à travers l'ouvrage « Acoustic theory of speech production », 1960) soit le seul qui vient des sciences physiques. Le fait de mettre en avant un objet désigné par un terme d'usage commun[3], la parole (*speech*), et non une discipline particulière (phonétique, phonologie, etc.), marque une évolution importante en autonomisant cet objet par rapport à des problématiques spécifiques qui pouvaient lui être associées jusque là, au moins dans l'univers académique. À travers ce jeu de désignation s'exprime un changement de statut de l'objet : la phonétique appartient aux phonéticiens, la parole appartient à tous. C'est un peu comme si, progressivement, les sciences de l'ingénieur s'étaient emparées d'un objet jusque là étudié principalement au sein des sciences de l'Homme, ce qui se perçoit aussi dans les évolutions du Conseil Permanent des Sciences Phonétiques (qui organise les congrès internationaux de Sciences Phonétiques) au sein duquel les chercheurs issus des sciences de l'ingénieur, absents jusqu'en 1961 prennent une place de plus en plus importante pour atteindre un tiers de l'effectif en 1991 (L.J. Boë et M. Grossetti, 1993). Cette évolution concerne aussi les problématiques, qui tendent à se focaliser de plus en plus sur des objectifs concrets comme le codage, la synthèse ou la reconnaissance, les approches de la parole en tant que produit spécifiquement humain passant parfois au second plan.

Toutes ces transformations ont pour cadre principal le monde anglophone, où les sciences de l'ingénieur se développent vigoureusement à partir des années quarante, en particulier dans les grandes universités technologiques américaines (MIT, Stanford, etc.) et les laboratoires industriels (Bell *Laboratories*). En France, la situation est très différente au lendemain de la guerre : les sciences de l'ingénieur sont très faiblement développées et la phonétique reste confinée dans les Facultés de Lettres sans grande activité expérimentale, à l'exception de quelques cas particuliers. Avant d'aborder l'analyse du processus d'émergence de la communauté « Sciences de la Parole », il est nécessaire de faire le point d'une part sur les évolutions de la phonétique et des sciences de l'ingénieur dans les universités au CNRS, et d'autre part sur les organisations de recherche technologique, industrielle ou publique[4] impliquées dans la naissance de la communauté, le Centre National d'Études des Télécommunications et le laboratoire d'IBM à La Gaude.

1.2 - La phonétique en France

La phonétique s'est développée au cours de la seconde moitié du XIX^{e} siècle autour d'objectifs liés à l'enseignement, à la lutte contre les pathologies affectant le langage, et à la connaissance des langues parlées. Elle trouve en France un bon terreau avec l'Association de Phonétique Internationale (1886) et surtout les travaux de l'abbé P. Rousselot, d'abord au sein de l'Institut Catholique de Paris, où il inaugure une chaire de Phonétique Expérimentale, puis dans le cadre l'Institut de

Phonétique de Paris, dont il est le premier directeur en 1911. Entre-temps a été créé l'Institut de Phonétique de Grenoble en 1904, essentiellement autour des questions d'enseignement du français aux étudiants étrangers attirés en grand nombre par l'Institut d'Électrotechnique. Un troisième Institut voit le jour en 1911 à Montpellier, sur la base d'un laboratoire universitaire créé en 1904.

Après l'essor du début du siècle, il semble que la phonétique expérimentale connaît, durant la période d'entre deux guerres, une phase de moindre activité, comme beaucoup d'autres sciences d'ailleurs. L'arrivée pendant l'occupation d'un phonéticien hongrois, G. Straka, élève de J. Chlumsky, jette les bases d'une réactivation de la recherche expérimentale en phonétique. G. Straka prend en 1950 la direction du nouvel Institut de Phonétique de Strasbourg et forme de nombreux élèves à l'approche expérimentale. Parallèlement à ce renouveau relatif, les instituts de phonétique commencent à se doter d'une instrumentation électro-mécanique ou électronique, à présent disponible sur le marché. L'arrivée d'une instrumentation moderne dans les instituts de phonétique pose au moins un problème de compétence technique : les phonéticiens ne sont pas au départ formés à l'utilisation, à l'adaptation, voire au développement de ces dispositifs et ont donc besoin de spécialistes lorsqu'ils désirent se livrer à des expérimentations. Au milieu des années cinquante, l'évolution interne de la phonétique, avec le renouveau de l'approche expérimentale et la mise en place d'équipements électroniques, est donc propice à l'établissement de coopérations entre phonéticiens et spécialistes des techniques, électroniciens en particulier. Nous verrons en détail comment des coopérations de ce type vont s'instaurer, à Grenoble en particulier. Signalons que trois nouveaux instituts de phonétique verront le jour par la suite : Aix-en-Provence (1960), Nancy (1970) et Besançon (1972).

1.3 - Électronique et informatique

On dispose de peu de travaux sur les débuts de l'électronique en France. Disons simplement que l'étude des courants faibles a peu à peu émergé au sein des laboratoires d'électrotechnique, notamment entre les deux guerres. En dehors de Paris et de l'École supérieure d'électricité, on trouve trois universités dotés de laboratoires de ce type, liés aux instituts d'électrotechnique créés au début du siècle : Grenoble, Toulouse et Nancy. Rappelons que ces instituts étaient des annexes des facultés des sciences, fondés grâce à des aides locales des milieux industriels ou des collectivités territoriales (Grelon, 1989). Si la chimie avait alors fait l'objet de nombreuses de créations d'instituts de ce type, les trois villes citées sont les seules où se soient développés des instituts consacrés à l'électricité. Or, cette discipline naissante devait attirer de très nombreux étudiants et se ramifier progressivement en trois sous-ensemble : l'électrotechnique elle-même (les courants forts), la radioélectricité puis l'électronique (les courants faibles) et l'hydraulique (autour des questions liées à l'hydroélectricité). L'Institut de Nancy, qui a connu quelques difficultés entre les deux guerres, n'a pas développé de laboratoire spécifique en radioélectricité comme ceux de Grenoble et Toulouse auxquels il faut ajouter l'Institut de radioélectricité de Bordeaux créé en 1920, mais peu actif jusqu'aux années soixante. C'est donc principalement à Paris, Grenoble et Toulouse que sont concentrées les forces de recherche et d'enseignement supérieur en électronique. Si à Paris l'électronique est dispersée entre différentes institutions (université, école d'électricité, instituts du CNRS, etc.), ce qui justifiera la création à la fin des années cinquante de l'Institut supérieur d'électronique de Paris, ce n'est pas le cas à Grenoble et Toulouse où elle se concentre à l'université et dans les écoles associées, les anciens instituts s'étant transformés en 1948 en Écoles nationales supérieures d'ingénieurs (ENSI). L'École nationale supérieure d'électronique et de radioélectricité de Grenoble (ENSERG) est alors une école à part entière. À Toulouse, le choix est fait de conserver les différentes filières de l'ancien institut d'électrotechnique au sein d'une même école, l'École nationale supérieure d'électrotechnique, électronique et hydraulique de Toulouse (ENSEEHT), mais l'organistaion de l'enseignement et de la recherche en électronique y est similaire à ce que l'on trouve à Grenoble.

L'informatique connaît des débuts difficiles en France avec les échecs subis dans la construction de calculateurs par les organismes de recherche (Rammuni, 1989, Mounier-Kuhn, 1993, et 1994 dans les mêms actes). Le CNRS avait fondé en 1946 à Paris l'Institut Blaise Pascal associant un laboratoire de calcul numérique dirigé par L. Couffignal et un laboratoire de calcul analogique dirigé par L. Malavard. Si le calcul analogique s'est développé dans de bonnes conditions à l'intérieur de ses limites, les tentatives pour développer de façon autonome des moyens de calcul numérique n'ont pas été un succès, ce qui fait que l'informatique telle que nous la connaissons aujourd'hui s'est vériatblement développée dans les institutions de recherche et d'enseignement supérieur à partir de l'arrivée sur le marché français des premièrs calculateurs construits autour d'une architecture de type Von Neumann, l'IBM 650, et le Bull Gamma ET, tous deux commercialisés en 1956.

Grenoble et Toulouse sont alors les universités les plus promptes à se saisir de ce domaine nouveau,

simultanément aux niveaux de la recherche avec la constitution de laboratoires en 1956 et 1957, des outils avec l'acquisition d'ordinateurs dès 1957, des enseignements universitaires avec la mise en place de certificats de calcul numérique et enfin dans la formation des ingénieurs avec la création de filières de formation spécifiques en 1958/59 (ENSI de Mathématiques Appliquées de Grenoble, ENSIMAG et filière de mathématiques appliquées de l'ENSEEHT, filière rebaptisée Informatique en 1967, l'école devant à cette occasion l'ENSEEIHT). Un des éléments expliquant le développement précoce de l'informatique à Grenoble en particulier est la pression exercée en faveur des enseignements de calcul numérique par les écoles d'ingénieurs, pression qui explique la présence de professeurs titulaires dans ce domaine, menant de front enseignement et recherche, éléments moteurs du développement des mathématiquees appliquées puis de l'informatique. Le même schéma vaut pour Nancy qui suivra l'exemple grenoblois après quelques années. C'est ainsi que jusqu'à la fin des années soixante, les principaux centres français de recherche et d'enseignement en informatique sont Paris, Grenoble, Toulouse et, à un moindre niveau, Nancy. Dans le cas de Grenoble, le poids de l'informatique est renforcé par l'installation en 1959 du Centre d'études pour la traduction automatique (le CETA), dirigé par B. Vauquois.

1.4 - La recherche technologique sur la parole

Deux organismes vont jouer un rôle important dans le domaine de la parole en cumulant chacun trois niveaux d'intervention : leurs recherches propres ; le financement incitatif de travaux d'équipes universitaires ou du CNRS ; la mise à disposition de données numérisées permettant aux chercheurs de mettre en œuvre leurs méthodes d'analyse.

Le Centre National d'Études des Télécommunications (CNET) est créé en 1944 sur la base de divers organismes existants (P. Bata, 1990). Le centre s'installe d'abord dans la région parisienne (à Issy-les-Moulineaux) avant de faire l'objet d'une premières grandes opérations d'aménagement du territoire qui conduit à la décision prise en 1954 de créer un grand établissement à Lannion. Le CNET connait alors une période de croissance forte, et l'ouverture en 1962 du centre de Lannion s'inscrit dans une phase d'évolution rapide des équipes comme des thèmes de recherche. C'est à Lannion que se sont développées les équipes amenées à s'intéresser à la parole, d'un côté dans le cadre de la commutation éléctronique et de l'informatique, de l'autre dans celui de l'acoustique. Nous verrons plus loin comment informaticiens et acousticiens seront amenés à travailler ensemble, mais il faut surtout insister à ce stade sur la mise au point par G. Ferrieu, au sein du département ETA (Études et techniques acoustiques), d'un vocodeur qui est opérationnel dès 1965.

Le centre scientifique implanté par IBM à La Gaude au début des années soixante est un peu un équivalent industriel du CNET de Lannion dans la mesure où une partie des travaux qui s'y développent a pour objet les télecommunications et la communication Homme-Machine. Une équipe du centre travaille sur les unités à réponse vocale, utilisant un IBM 7772 pour générer des échantillons numériques de parole.

2 - Logiques de convergence

La convergence vers des recherches expérimentales sur la parole utilisant une instrumentation moderne (électronique ou informatique), qui aboutit au colloque de 1967 et aux premières JEP, s'effectue progressivement entre 1955 et 1970 selon des logiques que l'on peut classer pour la clarté de l'exposé en deux grand types : les cas où l'intérêt pour ce type d'approche résulte de l'évolution interne d'une discipline et les cas où l'on assiste à une série d'interactions locales entre disciplines différentes.

2.1 - Démarches unilatérales

Au sein de cette catégorie, on peut distinguer plusieurs cas de figure selon le cadre discplinaire de départ des équipes amenées à travailler de façon expérimentale sur la parole à l'aide d'une instrumentation moderne.

Informaticiens à la recherche d'un objet.

À la fin des années cinquante, la reconnaissance des formes apparaît comme une direction importante de développement de ce qui ne s'appelle pas encore l'informatique. Les américains s'y essaient, d'abord au MIT et au sein d'IBM, puis, dès 1971, dans le cadre du projet ARPA (Advanced Research Projects Agency) SUR (Speech Understanding and Recognition). Les équipes françaises vont toutes se lancer dans cette voie à un moment ou un autre au cours des années soixante. À l'exception du cas de Grenoble sur lequel nous reviendrons, c'est par le jeu d'une spécialisation progressive à partir du problème générique de la reconnaissance des formes que la parole va être étudiée.
Les premiers à suivre cette voie sont les toulousains G. Perrenou et S. Castan qui commencent à travailler sur la reconnaissance des images par apprentissage en prenant pour base les travaux de F. Rosenblatt sur le *perceptron*. C'est à la suite d'un contact avec le CNET en 1965 qu'ils sont amenés à

s'intéresser à la parole. La connexion d'un vocodeur mis à disposition par le CNET comme organe d'entrées/sorties branché sur le calculateur IBM 7040 du Centre de calcul permet de tester sur des données de parole l'approche mise au point sur les images. Les premières JEP sont l'occasion pour l'équipe de Toulouse d'entrer en contact avec l'ensemble des chercheurs français travaillant dans le domaine.
Un scénario similaire se déroule à l'Institut de programmation de la Faculté des sciences de Paris où J.C. Simon enseigne la reconnaissance des formes au sein du DEA, avant d'encadrer plusieurs thèses dans ce domaine, la première étant celle de C. Roche qui applique des techniques de reconnaissance à des données issues, là encore, d'un vocodeur du CNET.
Le CNET sera aussi à l'origine de l'orientation vers la parole de chercheurs du Groupe d'Intelligence Artificielle (GIA) de Marseille : à la suite d'une présentation en 1973 d'un interpréteur portable du language Prolog à Lannion, P. Battani et H. Méloni commenceront à appliquer des techniques d'intelligence artificielle comme niveau supérieur d'analyse dans le cadre de la reconnaissance de la parole. Les données seront à nouveau fournies par le CNET.
La démarche au début des années soixante-dix de l'Institut de Mathématiques Appliquées de Grenoble (IMAG) avec la thèse de C. Bélissant sera du même type puisque, c'est à partir de travaux d'informatique en reconnaissance des formes que la parole sera étudiée (la numérisation des données étant effectuée sur un IBM 360).

Il est remarquable que le CNET soit souvent associé d'une façon ou d'une autre à ces évolutions, soit en tant qu'origine de l'incitation initiale soit pour fournir des données. Il faut bien convenir que cet organisme a contribué à amener vers l'objet parole une partie des informaticiens français.

On peut classer dans ce groupe les informaticiens de Nancy, même si le schéma est plus complexe. En effet, des recherches sur la parole sont déjà poursuivies par M. Lamotte, membre du Laboratoire d'électricité et d'automatisme, qui utilise un calculateur analogique. L'arrivée en 1969 d'un jeune agrégé de physique qui occupe un poste d'assistant en informatique, J.P. Haton, a pour effet d'orienter ces travaux vers l'utilisation d'un calculateur numérique (un T2000) associé à un dispositif de numérisation mis au point localement et d'un vocodeur CIT-Alcatel.

Phonéticiens à la recherche d'outils

Il ne semble pas que le laboratoire de phonétique de Nancy (devenu institut en 1970) soit à l'origine de cette évolution des électroniciens et informaticiens. L'influence réciproque semble aussi être faible, même si F. Lonchamp, l'un des membres de l'institut, commence en 1972 à utiliser des moyens numériques. Les deux équipes finiront par collaborer de façon régulière dans les années quatre-vingt.

Le dernier cas de logique unilatérale concerne l'Institut de phonétique d'Aix-en-Provence, dont le responsable, M. Rossi, s'intéresse aux questions d'intonation, soit un versant très expérimental de la phonétique. Il a effectué un stage au sein de l'institut de Grenoble, l'un des mieux équipés dans les années soixante. Par ailleurs, divers membres de l'institut d'Aix ont aussi participé en 1966 à un cours d'initiation à la programmation Fortran organisé par le centre IBM de La Gaude. Divers programmes seront réalisés pour les phonéticiens sur les machines du centre IBM. Il y a donc à Aix un intérêt fort pour l'expérimentation à partir d'instruments modernes (renforcé par une collaboration suivie avec les psychologues), intérêt qui aboutit au recrutement en 1969 d'un ingénieur électronicien, B. Teston, qui permettra à l'institut de suivre les évolutions de matériels, dont le passage au numérique n'est qu'une étape parmi d'autres.

2.2 - Interactions locales

En dehors du cas particulier du CNET, dont nous avons vu qu'il joue un rôle structurant pour les recherches sur la parole, notamment auprès des informaticiens et du centre IBM de La Gaude, dont nous verrons qu'il fournit des données numériques aux grenoblois, les collaborations impliquant des disciplines différentes s'effectuent toujours au sein d'un même pôle scientifique.

Le CNET-Lannion : une interaction prévisible qui aurait pu ne pas se produire

Le centre de Lannion voit se développer au cours des années soixante une équipe d'acoustique qui met au point un vocodeur et une équipe d'informatique. Il paraît donc assez naturel que l'idée de connecter le vocodeur sur un calculateur (ce qui est réalisé à Toulouse à l'incitation du CNET) aît fini par s'imposer. Il n'est pas certain que cela aurait pu s'effectuer si le principal concepteur du vocodeur (G. Ferrieux) et le responsable des calculateurs (J. Vincent-Carrefour) n'avaient pas partagé le même bureau... La connexion est opérée en 1967, permettant aux thèmes de la reconnaissance et de la synthèse de se développer.

Paris et le jeu des réseaux individuels

Au début des années soixante, dans le milieu des professionnels de la musique, de multiples recherches s'effectuent sur les structures physiques de la

musique, les rapports entre mathématiques et musique, etc. C'est dans cette ambiance culturelle et scientifique que se constitue le Groupe d'Acoustique Musicale (GAM) (que l'un de ses membres décrit comme une « *sorte de société secrète* », dans le cadre de la faculté des sciences de Paris, sous l'égide de R. Siestrunck (professeur de mécanique). Le groupe associe entre autres un chercheur CNRS (E. Leipp, auteur d'une thèse sur l'acoustique du violon), un professeur de musique (M. Castellengo) et un enseignant au conservatoire des arts et métiers (J.S. Liénard). Le GAM s'intéresse aux aspects musicaux de la voix et mène des recherche sur l'analyse et la synthèse de la parole en prenant pour base des tracés sonagraphiques. L'équipe met au point en 1965, avec l'aide du Laboratoire d'électronique de la faculté, une machine capable de synthétiser des sonagrammes simples, l'*Icophone*. R. Siestrunk met alors les membres du GAM en contact avec L. Malavard, directeur du Centre de Calcul Analogique (CCA) de l'Institut Blaise Pascal, qui propose au groupe d'utiliser l'ordinateur de son équipe pour piloter le synthétiseur. Une collaboration régulière se met alors en place entre le GAM et le CCA. Lorsque le CCA, transféré à Orsay, se transforme en Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI), il recrute deux membres du GAM et poursuit des recherches sur la parole, le GAM se recentrant sur la musique. C'est donc à travers l'acoustique musicale que la parole s'est dégagée comme objet de recherche pour les sciences de l'ingénieur, grâce à une série de relations individuelles (entre les membres du GAM au départ, entre Siestrunk et Leipp, entre Siestrunk et Malavard). Ce n'est pas là le seul cas d'interaction entre acoustique musicale et recherches sur la parole puisque l'on trouve des cas similaires au sein des laboratoires Bell ou du Royal Institute of Technology de Stockholm, ou même à Grenoble avec l'activité déployée par l'Association pour la Création et la Recherche d'Outils d'Expression (ACROE) au sein du LCP (Boë *et alii*, 1993).

Grenoble et la logique des institutions

Ce serait certainement très simplificateur de dire qu'à Grenoble, tous les ingrédients étaient rassemblés au sein d'un même ensemble universitaire et que leurs interactions ont fini par structurer un grand pôle de recherche sur la parole. La réalité est évidemment plus complexe, même s'il est certain que les points de départ et d'arrivée de cette version sont justes.

On peut faire débuter le processus en 1955 avec l'arrivée au sein de l'institut de phonétique d'un élève de G. Straka (dont nous avons vu plus haut qu'il est l'un des tenants de l'approche expérimentale en phonétique), R. Gsell. Dès son installation, ce dernier amorce une démarche tendant à associer la phonétique à la linguistique, au niveau des eneignements et des structures, avec l'introduction d'une nouvelle dénomination pour l'institut, qui devient alors l'Institut de phonétique et de linguistique de Grenoble (IPLG), dont il prend la direction en 1961. Pour la recherche, R. Gsell fait acquérir par l'institut une instrumentation moderne (magnétophones, sonagraphes, analyseur de spectre, etc.), ce qui justifiera le recutement de techniciens, de vacataires ingénieurs (J. Paillé), puis en 1967 d'un ingénieur électronicien, L.J. Boë. Parallèlement, il tente de tirer profit de l'environnement grenoblois en entrant en contact avec l'ENSERG (par le biais de son directeur, J. Benoit) et avec le CETA (B. Vauquois et l'un de ses collaborateurs, J.P. Tubach). B. Vauquois accepte de prendre la responsabilité d'un cours de linguistique appliquée, J.P. Tubach occupera durant un an un poste d'assistant à l'institut de phonétique. Divers membres de l'ENSERG font aussi des interventions dans le cursus de linguistique et de phonétique. Sur le plan de recherche, des collaborations se mettent en place sur le traitement statistique des sons du français et sur l'analyse instrumentale de l'intonation. R. Gsell propose aussi des thèmes de projets de fin d'études pour les élèves de l'ENSERG. Enfin, un spécialiste en otho-rhino-laryngologie de la faculté de médecine de Lyon (J.C. Lafon) fait des cours sur la physiologie de l'audition au sein de l'IPLG.

Cette recherche très large de coopérations locales menée par le directeur de l'IPLG a des conséquences importantes sur ses partenaires. C'est ainsi que la parole devient l'un des thèmes de recherche des équipes de l'ENSERG (en particulier celle qui se consacre à l'instrumentation et à la mesure des systèmes asservis, sous la responsabilité de R. Lancia), ce qui aboutit en 1960 à la création au sein de l'école du Laboratoire de la Communication Parlée (LCP). Travaillant au départ avec une instrumentation éléctronique, le laboratoire passera au numérique au début des années soixante-dix avec l'installation d'un PDP11/20. Au CETA, l'influence des phonéticiens se perçoit dans le travail de J.P. Tubach, à qui B. Vauquois propose en 1965 de de développer un système de reconnaissance de la parole qui pourrait servir d'interface pour la traduction automatique. Ce travail utilise des données fournies par le vocodeur d'IBM-La Gaude, ce qui s'inscrit dans une collaboration suivie entre les informaticiens grenoblois et IBM au cours des années soixante (le centre de calcul est équipé d'un IBM 360). Dans les années soixante-dix, IBM (qui a entretemps recruté J.P. Tubach) propose un sujet de thèse à une linguiste issue de l'IPLG, J. Vaissière, qui travaille à La Gaude sous la responsabilité scientifique de B. Vauquois.

3 - Les JEP

Les interactions entre disciplines sont donc très importantes à Grenoble, et il n'est pas surprenant de voir organisé en 1967 un colloque sur « les structures acoustiques de la parole » à l'initiative de R. Gsell (IPLG), J. Benoit (ENSERG), B. Vauquois (CETA) et de deux membres de l'Institut d'audiophonologie de la faculté de médecine de Lyon, P. Mounier-Kuhn et J.C. Lafon. Ce colloque est donc l'expression de la structuration d'un milieu de recherche sur la parole à Grenoble. Il attire 127 participants autour de 23 communications couvrant un champ très étendu : phonétique, linguistique, psychologie, physiologie, électronique, informatique, télécommunications, acoustique, acoustique musicale... On retrouve parmi les participants, en dehors des organisateurs et de leurs équipes, des représentants de la plupart des équipes citées plus haut[5], auquels s'adjoignent quelques participants étrangers (dont M. Wajskop de Bruxelles) et de nombreux universitaires lyonnais spécialistes d'audiophonologie ou de psychologie. Ce colloque est donc bien la première grande manifestation du domaine des sciences de la parole. Il est organisé par les responsables des institutions grenobloises (auxquels s'adjoingnent de manière un peu formelle les spécialistes lyonnais de l'audiophonologie) et se présente comme un colloque classique structuré en 6 séances présidées par les organisateurs ou des personnalités scientifiques nationales.

Pour les JEP de 1970 prévaudra une autre formule, beaucoup plus resserrée, avec une trentaine de participants et quatre conférences[6] : peu de thèmes, très ciblés, écartant délibérément une bonne partie des disciplines représentées en 1967. Le comité d'organisation est d'ailleurs totalement différent puisque l'on y trouve J.P. Tubach (CETA), R. Carré (ENSERG-LCP), M. Wajskop (directeur de l'institut de phonétique de Bruxelles, ancien assistant de l'IPLG), M. Rossi (Institut de Phonétique d'Aix-en-Provence), Mme Pela-Simon (Institut de Phonétique de Strasbourg) et R. Lancia (ENSERG). Aucun des initiateurs du colloque de 1967 n'est là. Les informaticiens et électroniciens, minoritaires à l'époque[7] sont à présent majoritaires, au côté des phonéticiens. Il y a donc eu entre les deux manifestation une réorganisation institutionnelle du champ, essentiellement au profit des chercheurs de terrain engagés dans des études expérimentales utilisant un appareillage électronique ou informatique, en même temps que disparaissaient les responsables institutionnels.

Entre les deux manifestations, il y a évidemment la bourrasque de 1968, particulièrement accentuée à Grenoble. L'IPLG est traversé à ce moment par une crise importante qui oppose le directeur et quelques proches collaborateurs aux membres les plus jeunes de l'équipe. Le conflit se solde par le départ du directeur et une réorganisation de l'institut (qui fusionnera en 1983 avec le LCP pour former l'Institut de la Communication Parlée). Cela explique l'absence des membres de l'institut dans l'organisation des premières JEP.

Un autre aspect des changements institutionnels du champ qui s'effectuent durant la période 1967 - 1970 est la création d'un groupe spécialisé au sein du GALF : le Groupe de Communication Parlée (GCP). L'idée semble s'imposer au cours d'une discussion informelle entre R. Carré et le président du GALF, à Tokyo (lors des sixièmes ICA). Le GCP organise les JEP en 1970 et devient le cadre associatif principal du nouveau domaine.

Conclusion

Il reste bien des choses à éclairer dans le processus qui a amené l'émergence de la communauté rassemblée régulièrement par les JEP à partir de 1970. La rupture opérée entre 1967 et 1970 a été constatée mais insuffisamment expliquée, faute d'éléments. De même, les logiques de convergence, mieux étudiées, laissent dans l'ombre des non convergences qui resteraient à approfondir : l'ensemble des laboratoires ou instituts de phonétiques qui ne s'associeront jamais à la réflexion menée dans les JEP, la faible présence au sein de la communauté de la physiologie et de la psychologie, etc.). Toutefois, cette rapide synthèse de la constitution de la communauté des sciences de la parole met en évidence plusieurs dimensions. Une dimension technologique d'abord : l'émergence de l'informatique a bouleversé profondément le champ, en introduisant de nouveaux acteurs et de nouvelles problématiques, tout comme l'instrumentation électro-mécanique avait pu le bouleverser aux États-Unis entre les deux guerres ; cette dimension se saisit aussi à travers l'importance du CNET et d'IBM, dont les moyens de numérisation se révèlent déterminants pour plusieurs équipes. Une dimension sociale, spatiale et institutionnelle qui se perçoit dans l'importance des interactions locales à l'œuvre dans certains sites. Les interactions locales, à Grenoble en particulier, sont en grande partie à l'origine de la structuration du domaine. Enfin, l'aspect interdisciplinaire de l'objet parole, évoqué rapidement dans cette synthèse, mériterait un approfondissement : les objets scientifiques impliquant autant de cadres disciplinaires différents ne sont pas si fréquents...

Références du texte

Bata P, 1990, «La naissance du CNET», *Le Centre National d'Études des Télécommunications. Naissance et genèse d'un centre public de recherche. 1944-1974*, CRCT

Boë LJ, 1990, *Éléments d'unification pour les Sciences de la Parole. Articulatoire, acoustique et structure du code,* Habilitation à diriger des recherches, INPG, Grenoble

Boë LJ, Grossetti M, Mounier-Kuhn PE, Cartier M, D'Alessandro C,Emerard F et Méloni H, 1993, «Informatique et Parole en France : un quart de siècle après la rencontre», 3° Colloque d'histoire de l'informatique en France, INRIA, Sophia-Antipolis

Boë LJ et Grossetti M, 1993, From analog to digital processing : a technological evolution in the study of speech», Colloque «Technological Change», Oxford, 7-9 Septembre

Boë LJ et Liénard JS, 1988, «La communication parlée est-elle une science ? En doutiez-vous ? Éléments de discussion et de réflexion suivis de repères chronologiques», XVII[e] JEP, GCP-SFA, pp. 79-92

Boë LJ et Liénard JS, 1988-1989, «Les sciences de la parole : constitution et développement. Éléments de discussion et de réflexion suivis de repères chronologiques», Bulletin de l'Institut de phonétique de Grenoble, Vol. 17-18, pp. 1-52

Fant G, 1960, *Acoustic theory of speech production. With calculations based on X-Ray studies or russian articulations*, Mouton, La Hague

Jakobson R, Fant G et Halle M, 1952, *Preliminaries to speech analysis*, MIT Press, Cambridge Mass.

Grelon A, «Les universités et la formation des ingénieurs en France (1870-1914)», *Formation et emploi*, n°27-28, 1989.

Mounier-Kuhn PE, 1993, «The failures of the French public research system in building pioneer computers», Colloque «Technological Change», Oxford, 7-9 Septembre

Rammuni G,1989, «La non construction du premier calculateur électronique au CNRS», Cahiers *pour l'histoire du CNRS*, n°4

[1] Cette communication s'inspire très largement d'un travail collectif réalisé dans le cadre de ce qui constitue actuellement la Commision « Parole et Informatique » du Groupe Francophone de la Communication Parlée et le GT0 du PRC informatique (Cf. L.J. Boë, M. Grossetti, P.E. Mounier-Kuhn, M. Cartier, C. d'Alessandro, F. Emerard et H. Méloni, «Informatique et Parole en France : un quart de siècle après la rencontre», 3e Colloque d'Histoire de l'Informatique en France, Sophia-Antipolis, Octobre 1993) ainsi que d'une réflexion menée avec L.J. Boë sur l'histoire des recherches concernant la parole (L.J. Boë et M. Grossetti, 1993). Ce qui est présenté ici constitue essentiellement un effort de synthèse de la part d'un sociologue travaillant sur les institutions scientifiques.

[2] Le terme « science de l'ingénieur » est utilisé ici par commodité pour désigner les disciplines des sciences de la matière actuellement regroupées dans le département du même nom du CNRS ou les départements d'*engineering* des universités américaines.

[3] Même si la parole était déjà constituée en objet scientifique à travers l'opposition opérée par Saussure entre la « langue » (le système) et la « parole » (matérialisation de la langue), opposition reprise en partie de W.D. Whitney.

[4] La physiologie et plus généralement les Sciences de la Vie sont moins impliqués dans le développement de la communauté, même si elles ont pu être associée au colloque de 1967, c'est pourquoi elles n'ont pas fait l'objet d'une analyse précise au sein du groupe et ne seront pas évoquées ici. Il reste certainement un travail à faire pour comprendre comment la dimension physiologique ou médicale ne s'est guère intégrée à la communauté alors en émergence malgré les intérêts scientifiques de ce domaine liés à l'objet parole.

[5] Entre autres : S. Castan du laboratoire d'informatique de Toulouse, R. Buron et A. Nemeth du centre IBM de La Gaude, M. Lamotte du LEA de Nancy, G. Ferrieu et M. Cartier du CNET, R. de Possel de l'Institut de programmation de Paris, M. Rossi de l'Institut de phonétique d'Aix-en-Provence, E. Leipp, M. Castellengo et JS Liénard des GAM.

[6] « La perception de la parole : orientations et perspectives » par M. Wajskop (Institut de Phonétique de Bruxelles), « Les appareils de synthèse et leurs applications » par R. Carré et J. Paillé (LCP - ENSERG), « La synthèse par règles de la parole », par A. Nemeth (IBM) et « Reconnaissance de mots et de langages parlés » par J.Y. Gresser (du CNET - Lannion).

[7] Environ un quart des participants selon un comptage effectué par L.J. Boë.

Apprentissage automatique pour l'interprétation sémantique

Renato De Mori

Université McGill
3480 rue Université, Montréal, Québec, Canada H3A 2A7

1 Résumé

Cet article décrit des techniques récentes pour la modélisation de l'interprétation automatique du langage avec une attention particulière aux modèles dynamiques et à l'apprentissage automatique.

2 Mots-clés

grammaire stochastique, interprétation sémantique, modèle de langage, modèles dynamiques, modélisation stochastique du langage.

3 Introduction

Actuellement, l'un des types populaires de système pour la reconnaissance automatique de la parole (RAP) consiste en deux composantes. Une composante acoustique fait correspondre des mots dans son vocabulaire à l'entrée acoustique, produisant une probabilité pour chacun des mots. La seconde composante, qui incorpore un modèle de langage, évalue pour chaque mot dans le vocabulaire sa probabilité d'apparition, étant donné une liste de mots hypothétiques élaborée précédemment. Chaque mot candidat sélectionné à l'origine par la composante acoustique est ainsi associé avec deux probabilités. La première est basée sur la ressemblance avec le signal observé et la seconde est basée sur la plausibilité linguistique de l'occurence de ce mot, immédiatement après les mots précédemment reconnus. La multiplication de ces deux probabilités produit une probabilité globale pour le mot candidat. Le problème de la compréhension d'une phrase est encore plus complexe. Une fois que le modèle stochastique que l'on vient d'introduire a été appliqué, il faut obtenir une interprétation à partir d'une séquence, ou de la collection des N meilleures séquences ou d'un réseau d'hypothèses de mots.

Cet article présente un aperçu des tendances récentes dans la modélisation stochastique du langage avec une attention particulière aux modèles dynamiques et à l'utilisation des grammaires stochastiques. Le problème de l'interprétation sémantique à partir de stratégies de classification apprises automatiquement est aussi discuté. Un apprentissage automatique est particulièrement important car la langue parlée ne respecte pas toujours des règles formelles complètement connues et les systèmes de reconnaissance introduisent des erreurs et des ambiguïtés.

4 Modèles dynamiques

Un modèle intéressant du langage incorpore un composant dynamique dans le calcul de la plausibilité linguistique. Supposons que la séquence de mots "le vieux ..." vient juste d'être reconnue, et que le mot "monsieur" suit ces deux mots dans 10% des cas dans le texte étudié, alors que le mot "manoir" suit cette séquence dans 1% des cas. Le modèle à trigrammes va assigner à "monsieur" la probabilité de 0.1 et à "manoir" la probabilité 0.01. Si la composante acoustique assigne à ces deux mots une probabilité a peu près égale, c'est le mot "monsieur" qui va être choisi. Pour une phrase isolée, cela serait un choix raisonnable à faire. Maintenant, supposons que plusieurs phrases précédentes contiennent le mot "manoir" et qu'aucune ne contienne le mot "monsieur". On devrait alors assigner une plus grande probabilité au mot "manoir".

Une forte évidence empirique existe en ce domaine. Des études sur trois collections de textes anglais et américains menées par S. Johansson montrent que "les fréquences des mots varient énormément selon le type du texte, et cela pour les mots de contenu ou les mots de fonction" [10,p34]. Un modèle de langage qui exploite les changements à court terme dans les fréquences d'utilisation de mots devrait fonctionner de façon plus performante que les modèles stochastiques courants du langage.

Un *modèle combiné* assigne une probabilité à chaque composant syntaxique (Part Of Speech: POS) tel que noms, adjectifs, verbes, etc., comme le modèle à trigrammes [9]. Pour un POS fixé, la prob-

abilité de n'importe quel mot lui appartenant est la moyenne pondérée de la fréquence du mot dans la catégorie de POS dans le Corpus d'apprentissage. A chaque POS est associée une mémoire tampon dite "cache". A un instant donné, durant la tâche de reconnaissance de la parole, le cache pour un POS donné va contenir les N derniers mots qui sont supposés avoir ce POS (N pouvant être égal à 200 par example). Si un mot est apparut fréquemment dans le passé récent, il va apparaître plusieurs fois dans le cache pour son POS (en supposant que, dans le but de l'explication, le mot a une seule possibilité de POS). Une plus grande probabilité va donc être assignée à ce mot.Si la fréquence d'apparition récente avait été basse, la probabilité aurait été plus basse. De cette façon l'inclusion d'une composante de cache satisfait le but de suivre dynamiquement les changements dans la fréquence d'utilisation des mots.

5 Traitement mathématique du Modèle Combiné

Le *modèle combiné* est maintenant introduit de façon mathématique. Rappelons que le modèle à trigrammes pur est basé sur les probabilités suivantes:

$$P(W_i = W|g_{i-2}, g_{i-1}) = \sum_{g_j \in G} P(W_i = W|g_i = g_j)P(g_i = g_j|g_{i-2}, g_{i-1}). \quad (1)$$

Le *modèle combiné* laisse la composante POS $P(g_i = g_j|g_{i-2}, g_{i-1})$ inchangée. La seule composante affecté est celle qui prédit la probabilité d'un mot étant donné le POS, $P(W_i = W|g_i = g_j)$. Dans le modèle à trigrammes, ceci est estimé par $f(W_i = W|g_i = g_j)$, calculé à partir du texte d'entraînement. Cela est certainement une bonne estimation de la moyenne autour de laquelle $P(W_i = W|g_i = g_j)$ fluctue. Cependant, cette quantité ne prend pas en compte la variance autour de cette moyenne.

Le passé récent est un bon indicateur de la direction de la variance. Ainsi, soit $C_j(W, i)$, l'estimation de la probabilité basée sur le cache pour le mot W à l'instant i pour le POS g_j. Ceci est calculé à partir des fréquences de W parmi les N plus récents mots appartenant au POS g_j. Le *modèle combiné* estime $P(W_i = W_j|g_i = g_j)$ par:

$$P(W_i = W|g_i = g_j) \simeq k_{M,j} \times f(W_i = W|g_i = g_j) + k_{C,j} \times C_j(W, i),$$

$$\text{avec } k_{M,j} + k_{C,j} = 1 \quad (2)$$

plutôt que par $f(W_i = W|g_i = g_j)$ uniquement. Ceci devrait permettre à l'estimation de $P(W_i = W|g_i = g_j)$ de dévier de sa valeur moyenne pour refléter temporairement des valeurs faibles ou basses. L'estimation des poids relatifs de $k_{M,j}$ et de $k_{C,j}$ est décrite en [11]. Ces valeurs obtenues ainsi maximisent la probabilité du texte d'étude. Notons que le *modèle à trigrammes* est simplement un cas particulier du *modèle combiné* obtenu en mettant tous les $k_{M,j}$ à 1.0 et tous les $k_{C,j}$ à 0.0. L'estimation de la composante POS $P(g_i = g_j|g_{i-2}, g_{i-1})$ pour le modèle combiné est réalisé de la même façon que décrite par Derouault et Mérialdo [7]. Pour s'assurer qu'aucun POS g_j n'a reçu une probabilité 0, un nombre arbitrairement petit 0.0001 a été ajouté selon l'approximation suivante:

$$P(g_i = g_j|g_{i-2}, g_{i-1}) \simeq l_1(g_{i-1})f(g_i = g_j|g_{i-2}, g_{i-1}) + l_2(g_{i-1})f(g_i = g_j|g_{i-1}) + 0.0001, \quad (3)$$

La description ci-dessus ignore la possibilité qu'un mot rencontré dans le texte échantillon puisse ne pas exister dans le vocabulaire (V) du système. Il y a plusieurs façons de calculer la probabilité qu'il va y avoir occurence d'un tel mot à l'instant i. Cette probabilité est estimée [11] par la formule de Turing qui utilise la fréquence d'un mot unique parmi tous les mots dans le texte étudié.

Un modèle *cache trigramme* peut aussi être considéré pour estimer $P(W_i = W| < W_1, W_2, ..., W_{i-1} >)$ de la façon suivante:

$$P(W_i = W| < W_1, W_2, ..., W_{i-1} >) \simeq a_1 \times \hat{P}(W_i = W|W_{i-2}, W_{i-1}) + a_2 \times [\sum g_j \in G \hat{P}(g_j|g_{i-2}, g_{i-1}) \times [k_{M,j} f(W|g_j) + k_{C,j} C_j(W, i)]], \quad (4)$$

$$a_1 + a_2 = 1, k_{M,j} + k_{C,j} = 1.$$

Différents types de cache à unigrammes, bigrammes, trigrammes ont été évalués par plusieurs chercheurs avec des corpora de plusieurs langages. Les résultats sont décrits en [13] et [14].

Cette idée a été incorporée dans une théorie plus générale visant à calculer les probabilités du modèle du langage à partir de plusieurs sources de connaissances produisant des composants au calcul de la probabilité d'une séquence de mots [17].

6 Modélisation de langage utilisant des grammaires hors-contextes stochastiques

La compréhension automatique de la parole (CAP) diffère de la reconnaissance automatique de la parole parce qu'elle doit produire une représentation

sémantique d'un message parlé plutôt qu'une séquence de mots reconnus. La structure d'une représentation sémantique dépend de l'utilisation qui doit en être faite. Dans la plupart des applications CAP, une phrase est prononcée ou un dialogue s'amorce pour produire certaines actions. Des examples d'action basées sur une représentation sémantique sont des accès à une base de données.

Les interprétations dépendent souvent des composantes de l'arbre syntaxique d'une phrase. Souvent, il n'est pas nécessaire d'avoir un arbre syntaxique complet.

La section suivante introduira une méthode pour caractériser l'information nécessaire dans une phrase et pour produire une représentation conceptuelle. On verra que, au moins dans certains domaines d'application, une connaissance complète de la structure d'une phrase n'est pas nécessaire. Il est alors important de développer des méthodes de calcul de la probabilité d'une phrase même si l'on ne dispose pas d'une analyse syntaxique complète. Des résultats sur le calcul des probabilités dans ces cas, utilisant des grammaires stochastiques, sont décrits en [3] et [9].

7 L'approche sémantique globale pour la CAP

Il est possible de distinguer deux types d'approches principales pour la compréhension des langages parlés. Une approche tente de produire une analyse syntaxique complète qui soit valable pour tous les mots dans la prononciation d'une phrase, alors que les autres comptent sur des mots-clés relevés et des composants de phrases, et utilisent des contraintes sémantiques globales pour arriver à une représentation conceptuelle. Toutes ces approches utilisent la syntaxe à un niveau local, pour traiter les phrases nominales, prépositionnelles, etc... Mais elles diffèrent au niveau global, là où l'approche sémantique globale a pour but de remplir un gabarit approprié sans fournir une analyse globale. L'approche purement syntaxique tend à se décomposer lors de la présence de mots inconnus, de nouvelles constructions linguistiques, d'erreurs dans la reconnaissance, et de phénomènes spontanés dans la parole comme les répétitions. Ces évènements causent moins de problèmes pour l'approche sémantique globale. Celle-ci peut "éviter" les séquences de mots problématiques à l'intérieur de la prononciation d'une phrase.

Récemment, de nombreux groupes travaillant dans le reconnaissance de la parole ont été attirés par l'approche sémantique globale. Un bon example est fourni par les groupes travaillant sur le système d'information d'une agence de voyages (ATIS Air Travel Information System), (cf [1] et [18]). En effet, les chercheurs qui ont commencé avec une approche purement syntaxique ont tendance à incorporer des éléments de l'approche sémantique globale, alors que l'on observe très peu de mouvements dans le sens inverse. Par example, le Template Matcher du groupe SRI [1] trouve et analyse des phrases clés, génère une liste des endroits clés possibles que la phrase pourrait remplir. Il choisit alors un gabarit général basé, en partie sur la maximisation du nombre de composants qui pourraient convenir pour ces endroits clés dans ce gabarit, et en partie sur la présence de certaines autres mots-clés en corrélation avec un gabarit particulier. Bien que certains affinements et extensions aient été faites, le Template Matcher fonctionne toujours de la même façon [1]. L'on peut dire la même chose de l'approche par décodage statistique d'ATT [12].

Par contraste, le groupe du MIT [15], et le groupe BBN [16], ont récemment implémenté des analyseurs basés sur la sémantique globale qui agissent comme des sauvegardes quand l'analyseur original purement syntaxique échoue. Le groupe BBN a même trouvé que les meilleurs résultats étaient obtenus quand l'analyseur sémantique de sauvegarde fonctionnait tout seul, sans aucune analyse préliminaire effectuée par l'analyseur original purement syntaxique.

Ainsi, un consensus en faveur de l'approche sémantique semble s'être développée au sein de la communauté de la compréhension de la parole, et cela même parmi ces groupes qui, à l'origine, étaient plus en faveur de l'approche purement syntaxique. Dans l'approche sémantique globale, la compréhension dépend de la reconnaissance d'ilôts de mots au sein de la phrase prononcée. Ces ilôts de mots possèdent l'une des trois fonctions suivantes :

- 1. Ce sont de potentiels "remplisseurs" d'endroits clés qui doivent être analysés. Par example, ils donnent une heure, la date, ou un lieu qui va remplir l'endroit clé dans la représentation conceptuelle.
- 2. Ils aident à indiquer la catégorie sémantique d'une phrase prononcée - i.e. ils identifient le "format" ou le "gabarit"-. Dans certains systèmes, il peut aussi être important d'identifier des catégories illocutionnaires: est-ce qu'il s'agit d'une question ou d'une commande?
- 3. Ils indiquent à quel endroit clé tel remplisseur appartient. Par example, dans ATIS, il n'est pas suffisant de reconnaître qu'il s'agit d'une heure et transformer cela dans une forme standard, il faut aussi savoir si cela s'applique à l' arrivée d'un vol ou à un départ. De façon similaire, un nom de ville peut représenter une origine, une

escale, ou une destination. Des mots tels que "vers" "de" adjacents au remplisseur peuvent fournir les indices nécessaires, mais, souvent, les mots-clés gouvernant le remplissage des endroits clé sont plus loin et moins clair.

En général, les analyseurs syntaxiques locaux sont écrits pour identifier des ilôts de mots constituant des remplisseurs potentiels. Les deux autres types d'ilôts de mots sémantiquement important sont beaucoup plus difficiles à identifier à l'avance, parce qu'ils n'apparaissent pas directement dans la représentation conceptuelle. Par example, de subtiles différences dans les termes employés, et cela dans des parties discontinues de la phrase prononcée, peuvent mener à une distinction entre deux formats.

Des arbres de classification de chaines (SCTs: String Classification Trees) [12] ont été conçues pour identifier automatiquement des ilôts de mots qui comportent de l'information sur le format approprié pour une phrase prononcée, ou encore sur l'endroit clé approprié pour un remplisseur reconnu précédemment. Ils n'éliminent pas la nécessité d'une analyse locale des remplisseurs - typiquement des phrases nominales ou des phrases propositionnelles - mais ils éliminent le besoin de coder à la main la majorité du travail fait par un analyseur sémantique global. Par example, le Template Matcher du SRI contient un grand nombre de règles codées à la main pour assigner à la totalité de la phrase prononcée un gabarit unique. Les deux types de règles peuvent être remplacées par des arbres de classification entraînés par des données étiquetées de façon appropriée. Même si l'on prend en compte la durée de l'étiquetage, construire des analyseurs sémantiques globaux en augmentant les arbres des classification sera plus rapide et plus préçis que d'écrire des règles codées à la main.

L'un des avantages des arbres de classification, qui fait d'eux un outil utile pour développer des systèmes de CAP, est le fait qu'ils compensent partiellement les données bruyantes. Quand les arbres de classification sont employés sur des séquences de mots, des hypothèses sont générées par un système de compréhension de la parole et classées par la suite. Les mots-clés choisis vont tendre à être des mots à taux de non reconnaissance peu élevé. Alors que les arbres de classification ont à l'origine été développés pour la tâche ATIS, les algorithmes et logiciels sont complètement généraux: ils peuvent être utilisés pour apprendre des mots-clés sémantiquement importants à partir de séquences de mots dans n'importe quel contexte.

Les algorithmes d'amélioration des arbres de classification vont maintenant être décrits, et des résultats expérimentaux vont être donnés.

8 Apprentissage des arbres de classification

Un arbre de classification binaire est un arbre binaire dont chacun des noeuds est associé à une question oui-non, un sous-arbre OUI et un sous-arbre NON, et dont les feuilles de noeuds correspond à une catégorie. Pour classer une donnée, il faut appliquer la question à la racine et suivre le sous-arbre OUI ou NON selon la réponse à la question. Il faut appliquer cela récursivement jusqu'à ce que la donnée parvienne à un noeud feuille, dont la catégorie va être assignée à la donnée.

Des algorithmes efficaces pour l'entraînement des arbres de classification binaires à partir de données d'étude étiquetées ont été développés. Beaucoup d'entre eux sont contenus dans le livre de Breiman, Friedman et Al [2].Deux importantes et récentes contributions sont décrites en [4] et [8].

Pour implémenter les algorithmes d'entraînement d'arbres, l'on doit fournir trois éléments:

- 1. L'ensemble des questions oui-non possibles et qui peuvent être appliquées aux données;
- 2. Une règle pour sélectionner la meilleure question à chaque noeud sur la base des données étudiées;
- 3. Une méthode pour "tailler" les arbres afin d'éviter le "sur-apprentissage"

L'on doit alors se procurer des algorithmes avec un Corpus d'étude comportant des données pré-classées. Ces algorithmes vont générer un arbre qui pourra classer les nouvelles données.

Un problème intéressant est la manière dont l'ensemble des questions possibles (qui est beaucoup plus grand que le sous-ensemble finalement sélectionné pour remplir les noeuds) est généré. Pour choisir une question dans cet ensemble, nous utilisons le critère de Gini (cf [2]). Pour éviter le surentraînement des arbres, un algorithme a été décrit dans [8].

Considérons un ensemble de chaînes de caractères, chacune consistant en une séquence de symboles d'un alphabet. Chaque chaîne appartient à une catégorie. Pour nos examples, nous allons supposer que l'alphabet est l'ensemble des lettres minuscules a,b,...,z, et que les catégories sont des entiers dans l'ensemble 1,2,...,N. Nous aurons également besoin d'un symbole "wild-card" ou "pas important": +. Ce + est équivalent à une séquence inconnue de symboles, de longueur au moins un. Dans les applications des langages naturels, les symboles seront des mots du vocabulaire (c'est à dire que le lexique constitue l'alphabet).

Une chaîne vient juste de parvenir au noeud racine d'un arbre de classification: quelle question oui-non pourrait-on poser à son propos? Supposons que la chaîne de caractères comporte au moins un symbole. Puisque l'on ne sait rien d'autre au propos de la chaîne, celle-ci est de la forme connue +. A tous les noeuds, les règles de génération de questions impliquent le remplacement d'un + dans la forme connue par une chaîne de caractères. Si X est une variable pouvant représenter n'importe quelle lettre de l'alphabet, + peut être remplacé par X, par X+, par +X, et par X+X. Au noeud racine, où les formes connues sont décrites par +, ces règles fonctionnent pour produire les questions types suivantes:

- 1. Questions du type: "Est-ce que la forme de la chaîne est X?" (26 questions individuelles, comme X peut être un symbole de a à z)
- 2. Questions du type: "Est-ce que la forme de la chaîne est X+?" (26 questions)
- 3. Questions du type: "Est-ce que la forme de la chaîne est +X?" (26 questions)
- 2. Questions du type: "Est-ce que la forme de la chaîne est +X+?" (26 questions)

Le critère de Gini est utilisé sur les données d'étude étiquetées pour sélectionner l'une de ces questions afin d'occuper le noeud racine. Supposons qu'il s'agisse du type de question 4, "Est-ce que la forme de la chaîne est +e+?" Les chaînes dans la donnée d'étude qui vont mener à une réponse OUI à cette question seront passées au noeud fils OUI, celles menant à une réponse NON seront passées au noeud fils NON. Par example, la chaîne "xyebeq" irait à OUI tandis que la chaîne "rvk" irait à NON.

Dans le noeud OUI, un ensemble des questions possibles quelque peu différent est généré au moyen des mêmes règles. Rappelons que chaque + dans la forme connue peut être remplacée par X, X+, +X, +X+, et nous spécifierons qu'un seul remplacement est permis par question. Ainsi, à partir de la forme connue +e+ pour les chaînes dans le noeud sous-fils OUI, l'on obtiendrait 9 questions types. 4 des questions seraient obtenues par remplacement du premier +, 4 seraient obtenues par remplacement du second +, et l'une concernerait la longueur. Par example, des questions comme: "Est-ce que la forme de la chaîne +e est +X?" (26 questions, comme X est affecté à a,b,...,z) sont obtenues en remplaçant le second + dans +e+ par +X.

Pour tous les noeuds dans l'arbre, les questions pour leus noeud fils OUI sont générées exactement de la même façon, c'est à dire en remplaçant le +s dans la forme connue. Les questions considérées dans le noeud fils NON d'un noeud sont les mêmes que celles considérées dans le noeud parent, hormis le fait que la question posée pour le parent n'est pas reposée. Parmi les questions considérées à un noeud, celle que satisfait le mieux le critère de Gini est choisie pour remplir le noeud.

En générant les questions de cette façon, l'on obtient un arbre de classification à symbole simple. Chaque noeud feuille correspond à un format de symboles connus et de blancs dans lesquels les symboles connus sont ceux qui sont les plus importants pour déterminer la classe de la chaîne. Si nous entraînons l'arbre de classification avec les phrases étiquetées dans lesquelles les symboles sont des mots w_i dans le lexique, les noeuds feuille représentent des formats de mots-clés. Par example, une feuille pourrait affecter à une classe toutes les phrases de la forme w_j+ (phrases commençant avec le mot w_j), alors qu'une autre pourrait affecter à une autre classe toutes les phrases de la forme $+w_2w_4+$ (phrases contenant la séquence de mots w_2w_4).

L'apprentissage peut être fait à partir d'une traduction d'un Corpus de phrases dans le langage d'interrogation de bases de données SQL fourni par ARPA.

La majorité des phrases du premier ensemble d'apprentissage peut être affectée à l'une des six structures. Ces schémas ne sont pas subjectifs, ils sont définis en terme de la translation SQL - suppléée par ARPA- de chaque phrase. Ainsi, nous pouvons fournir les algorithmes d'entraînement des arbres de classification avec des transcripts étiquetés et avec le numéro de schéma correspondant, et alors voir combien de fois l'arbre de classification résultant classifie correctement un ensemble disjoint de transcripts. Les schémas sont numérotés de la façon suivante:

- **1. AIRCRAFT**
- **2. FARE**
- **3. FLIGHT**
- **4. FLIGHT-FARE**
- **5. GROUND**
- **6. MEANING.**

Pour ces expériences, les algorithmes d'entraînement des arbres de classification ont reçu des phrases qui n'avaient pas été préanalysées de quelque manière que ce soit (hormis l'étiquetage avec le numéro de schéma).

9 Entraînement des arbres de classification d'ensembles

L'arbre de classification entraîné à partir des données étiquetées ATIS s'est révélé être extrêmement dis-

proportionné, avec la plupart des sous-arbres NON contenant beaucoup plus de noeuds que leurs noeuds frères OUI. Ceci s'explique par le fait que les phrases contiennent beaucoup moins de mots que le lexique anglais. Donc, si l'on demande si une phrase contient tel ou tel mot, la probabilité que la réponse soit "NON" est supérieure à 0.5. Pendant l'entraînement de l'arbre de classification à symbole simple, la plupart des examples étiquetés arrivant à une question donnée va mener à la réponse "NON", d'où la disproportion de l'arbre.

Cela reflète un véritable problème des arbres de classification à symbole simple . Les questions dans les noeuds de ces arbres sont incapables de regrouper des mots quasiment synonymes apparaissant aux mêmes positions dans les phrases. Par example, le mot "taxi", "limo", ou "bus" apparaissant au milieu d'une phrase est un bon indicateur du schéma GROUND. Mais un arbre de classification à symbole simple ne peut pas contenir la question: "Est-ce que la forme de la phrase est +S+, avec S symbolisant un élément ou plus de l'ensemble *taxi,limo,bus*?"

Pour remédier à ce problème, nous avons conçus des algorithmes pour entraîner des arbres ayant des ensembles dans les formes des noeuds (SCT: Set Classification Trees). Dans les SCTs, les noeuds peuvent contenir des questions sur des ensembles de mots dans une certaine position. Une question typique dans un arbre "set-membership" pourrait être: "est-ce que la forme de la phrase est $+S_1 + S_2+$?" avec S_1, S_2 des ensembles de mots. Une chaîne de la forme +u+v+, avec u dans S_1 et v dans S_2, mènerait à la réponse OUI dans ce cas. En d'autres termes, nous demandons seulement qu'au moins un mot de chaque ensemble apparaîsse à la bonne position.

Nous allons éviter de rentrer dans le détail des algorithmes qui génèrent des questions pour les arbres de classification "set-membership", qui sont plus compliqués que ceux qui génèrent les questions pour les arbres de classification à symbole simple. Notons simplement qu'à chaque position dans la forme connue au niveau d'un noeud donné, un certain nombre d'ensembles de mots, exponentiel par la taille du lexique, pourrait être généré. Pour éviter cela, l'algorithme qui génère les ensembles de mots possibles à une certaine position emploie une heuristique qui commence par la meilleure question à symbole simple à cette position et essaie alors de l'améliorer de façon incrémentale en ajoutant un symbole à la fois. Finalement, notons qu'un ensemble de mots obtenus près du noeud racine peut être divisé par une question dans un noeud descendant. Par example, si une question posée plus tôt a établi que les phrases dans un noeud sont de la forme +taxi,limo,bus+, l'arbre peut choisir une question qui distinguera entre ces modes de transport, tel que: "Est-ce que la forme est +bus+?".

Les algorithmes d'entraînement d'arbres d'ensembles aboutissent à un arbre à 45 noeuds sur les données étudiées décrites dans la section précédente. Les algorithmes "set-membership" génèrent souvent des ensembles de mots qui correspondent à des classes de mots sémantiques. Une question candidate qui fût formulée mais rejetée en fin de compte, en la faveur d'une autre meilleure, contient l'ensemble de mots: *define, what, describe, explain.*

D'autres ensembles de mots ayant été réduits contenaient une longue liste de numéros de vols ou de noms de ville. Ceci est très encourageant car - sans aucune intervention de lêtre humain - ces algorithmes qui entraînent des arbres d'ensemble ont presque "inventé" des classes sensibles de mots. Ceci suggère que des arbres de meilleure qualité encore pourrait être obtenus en créant des algorithmes possédant une certaine connaissance *a priori*.

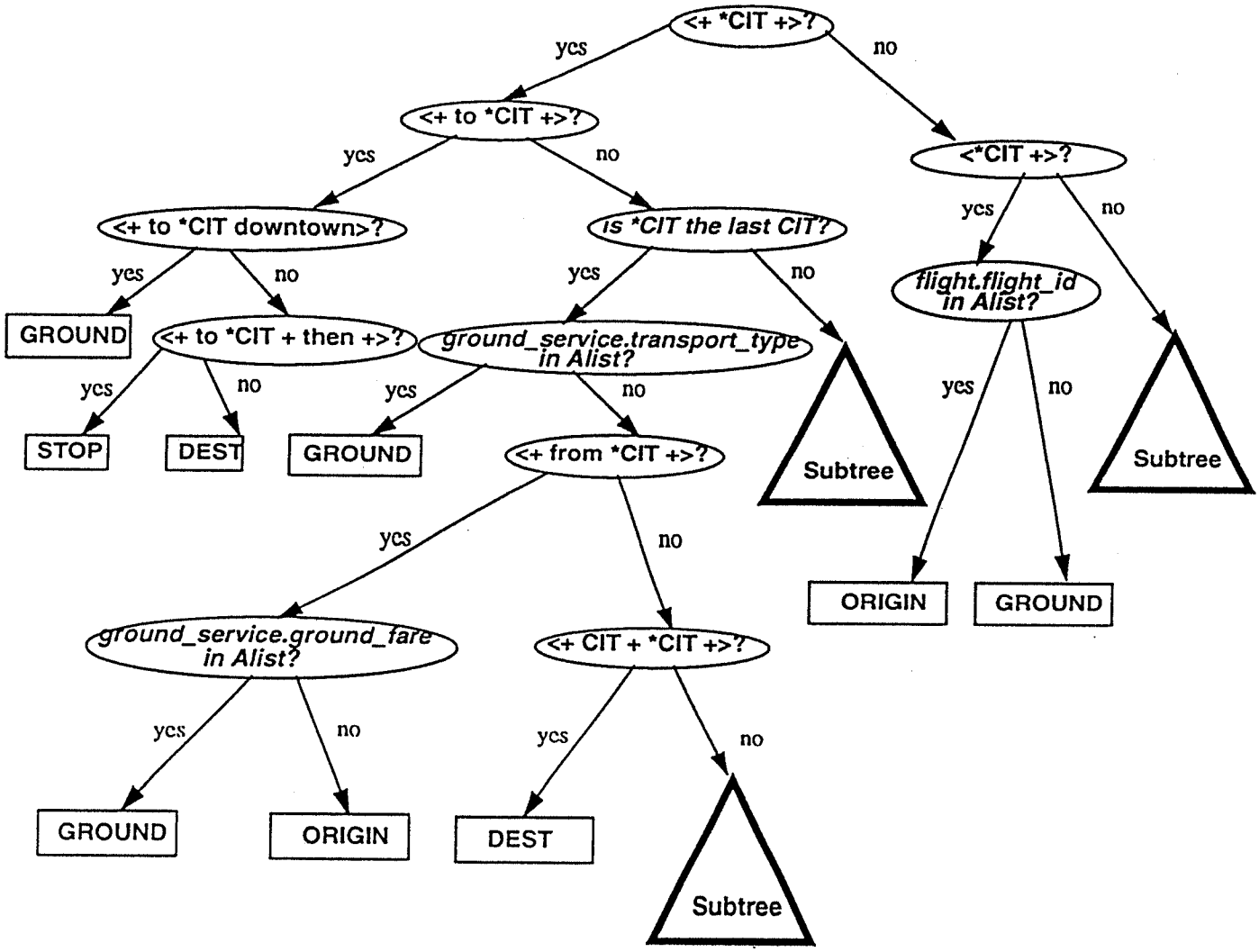
10 Améliorations du système de classification

Des séquences de mots sont d'abord soumises à un analyseur syntaxique de type "chart-parser" qui identifie les noms sémantiquement importants et les groupes propositionnels tels que les noms de ville, les expressions temporelles, les codes des vols, les noms des lignes aériennes, etc., et les remplace avec des symboles génériques. Les valeurs correspondantes sont conservées pour un usage ultérieur. Par example, avec les symboles CIT et TIM remplaçant respectivement les noms de ville et les heures, la phrase "show flights from Boston to New York and then on to Washington before noon" deviendrait "show flights from CIT to CIT and then on to CIT TIM". La séquence partiellement analysée résultante est ensuite passée à travers une forêt "d'Arbres de Classification Sémantiques". Chaque arbre est responsable de la géneration d'un aspect différent de la représentation sémantique. Ces arbres de classification sont entraînés automatiquement à partir de données d'étude étiquetées.

Les améliorations des arbres de classification sont basées sur l'idée d'une interdépendance entre les nombreux arbres qui génèrent ensemble la représentation sémantique à partir des hypothèses de séquences de mots. Les arbres de classification peuvent dorénavant poser des questions qui dépendent du résultat en sortie d'autres arbres de classification.

La figure 1 montre la sous-chaîne d'arbres de classification entraînée pour classer les CITs.

Comme le symbole CIT peut apparaître plusieurs fois dans une phrase, notons "*CIT" le CIT courant que l'arbre de classification de la figure 1 essaie de classifier. Une phrase avec N CITs va passer dans l'arbre N fois, avec "*" marquant un CIT différent



Alist = *(ground_service.city_code, ground_service.airport_code, ground_service.transport_type)*
"is there ground transportation from the airport to CIT => GROUND downtown"

Alist = *(flight.flight_id)*
"show flights from CIT => ORIGIN to CIT=>STOP and then on to CIT=>DEST before TIM"

Alist = *(flight.departure_time, flight.flight_id)*
"CIT=>ORIGIN to CIT=>DEST with all flight times please"

Alist = *(ground_service.ground_fare)*
"how much is a taxi from CIT=>GROUND to the airport"

Figure 1: SCT pour CIT

à chaque fois. Considérons "show flights from CIT to CIT and then on to CIT TIM". Pour chercher le rôle du premier CIT, "show flights from *CIT to CIT and then on to CIT TIM" est soumise à l'arbre. Ceci correspond à la forme < + * CIT +> (+ symbolise une forme inconnue) et donc la réponse à la question à la racine est "yes". Puisque "*CIT" n'est pas précédé par "to", la question suivante mène à "no", comme la réponse à la question de savoir si ce CIT est le dernier CIT de la phrase. Ensuite, la phrase est envoyée profondément dans le large sous-arbre représenté par un triangle. Le premier CIT va se voir affecter un rôle (espérons qu'il s'agisse de son rôle correct: ORIGIN). Pour classer le deuxième CIT, "show flights from CIT to *CIT and then on to CIT TIM" est soumise à l'arbre de classification. Les correspondances avec les formes < + * CIT +> et < +to * CIT+ > sont établies mais pas avec la forme < +to * CITdowntown>. Ensuite, la forme < +to * CIT + then + > est reconnue afin de pouvoir affecter au second CIT le rôle correct STOP (stop-over city). Finalement, "show flights from CIT to CIT and then on to *CIT TIM" va mener au rôle correct DEST pour le troisième CIT.

La partie basse de la figure 1 montre quelques examples qui ont servi à entraîner cet arbre de classification. Pour le moment, il y a un arbre de classification par attribut, et chaque arbre décide indépendamment si un attribut doit apparaître ou non dans la liste. Cependant, si l'on spécifie un ordre de génération des arbres attributs, l'on pourrait permettre aux derniers arbres de poser des questions sur les décisions prises par les premiers arbres. Cela assurerait que des attributs mutuellement incompatibles n'apparaîssent jamais sur la même liste d'attributs. De la même façon que pour les sous-chaînes d'arbres à attributs dépendants, c'est un moyen pour assurer la cohérence interne de la représentation sémantique.

Les résultats expérimentaux de la compétition ARPA de décembre 1993 sont décrits dans la table 1. Cette table donne les pourcentages d'erreur d'interprétation des phrases du test du système ATIS de type A. Pour plus de détails sur le Corpus, il faut se référer à [20].

Système	Erreur
ATT	7.4
BBN1	9.6
BBN2	16.1
CHANEL	12.3
CMU	6.0
MIT-LCS	10.0
SICSA	14.7
SRI	14.3
UNISYS	28.6

Table 1

Il y a deux systèmes qui sont basés sur un apprentissage automatique: CHANEL et BBN2. On peut voir que CHANEL a de meilleures performances que certains systèmes entraînés manuellement.

11 Conclusion

Nous avons montré comment les arbres de classification peuvent apprendre automatiquement des règles qui traitent de deux problèmes rencontrés avec les analyseurs sémantiques globaux dans les systèmes de compréhension:

Un module langage naturel basé sur les arbres de classification a été construit pour la tâche ATIS et a pu être évalué lors de la compétition de décembre 1993 organisée par l'Advanced Research Project Agency (ARPA) aux Etats Unis.

12 Remerciements

Un vif remerciement à Anna Corazza, Roland Kuhn et Sandrine Le Calvez. La recherche décrite dans cet article s'inscrit dans le cadre des activités de l'Institut pour la Robotique et les Systèmes Intelligents, un réseau canadien de centres d'excellence.

13 BIBLIOGRAPHIE

1. D. Appelt and E. Jakson. "SRI International February 1992 ATIS Benchmark Test Results" en [5]

2. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*, Wadsworth Inc., 1984.

3. A. Corazza, R. De Mori, R. Gretter, and G. Satta. "Language modeling using stochastic context-free grammars", *Trans. IEEE PAMI*, V. 13, no. 9, pages 163-170, 1993.

4. P. Chou. "Optimal Partitioning for Classification and Regression Trees", *Trans. IEEE PAMI*, V. 13, no. 4, pages 340-354, Apr. 1991.

5. *Proceedings of the Fifth DARPA Workshop on Speech and Natural Language*, Arden Conference Centre, Harriman, NY, Feb. 1992. Morgan Kaufmann Inc.

6. *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, Feb. 1991. Morgan Kaufmann Inc.

7. A. M. Derouault and B. Mérialdo "Natural language modeling for phoneme-to-text transcription." *IEEE Trans. Pattern Anal. Machine Intell.*, vol.PAMI 8 pages 742-749,Nov. 1986.

8. S. Gelfand, C. Ravishankar, and E. Delp. "An Iterative Growing and Pruning Algorithm for Classification Tree Design",

   Trans. IEEE PAMI, V. 13, no. 2, pages 163-174, Feb. 1991.

9. F. Jelinek, J. D. Lafferty, and R. L. Mercer, "Basic Methods of Probabilistic Context Free Grammars," in *Speech Recognition and Understanding* (R.D. Mori and P.Laface, eds.), Berlin, Germany: Springer-Verlag, 1992.

10. S. Johansson, E. Atwell, R. Garside, and G. Leech *The tagged LOB Corpus User's Manuel* Norwegian Computing Centre for the Humanities, Bergen, Norway, 1986.

11. R. Kuhn and R. De Mori, "A Cache-Based Natural Langage Model for Speech Recognition" *IEEE Trans. Pattern Anal. Machine Intell.* Vol.12 No.6, pages 570 à 583 June 1990

12. R. Kuhn and R. De Mori, "Learning speech semantics with keyboard classification." *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 55-58. Minneapolis, MN, 1993.

13. R. Lau, R. Rosenfeld, and S. Roukos. "Trigger-based language models: A maximum entropy approach." *In Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 45-48. Minneapolis, Minnesota, USA, 1993.

14. N. Ney, U. Essen, and R. Kneser."On structuring probabilistic dependencies in stochastic language modeling." *Computer, Speech and Language*, 7:101-138, 1993

15. J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, vol. I. Addison-Wesley Co., 1984.

16. R. Pieraccini, E. Tzoukermann, et al. "Stochastic Representation of Conceptual Structure in the ATIS Task", en [6], pages 121-124.

17. R. Rosenfeld. "A hybrid approach to adaptative statistical language modeling." pages 76-81. Princeton, NJ, March, 1994.

18. S. Seneff. "Robust Parsing for Spoken Language Systems", *ICASSP 92*, V. I, pages 189-192, San Francisco, CA, Mar. 1992.

19. D. Stallard and R. Bobrow. "Fragment Processing in the DELPHI System" en [6].

20. *Proceedings of DARPA Speech and Natural Language Workshop* Princeton, NJ, March, 1994.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O1

HISTOIRE DES SCIENCES DE LA PAROLE

LA RECHERCHE PUBLIQUE FRANÇAISE ET LES PREMIERS ORDINATEURS : UN ECHEC LOURD DE CONSEQUENCES (1939-1959)

P.-E. MOUNIER-KUHN

CNRS et Centre Roland-Mousnier - Histoire et Civilisation, Paris-IV Sorbonne

De tous les pays industrialisés, la France est le seul où la recherche publique n'a construit aucun ordinateur à l'époque pionnière — avant la fin des années cinquante. On peut expliquer cet échec, mettant en cause divers laboratoires, par l'impact de la défaite de 1940, par la situation difficile des mathématiques appliquées dans le milieu universitaire de l'époque, par les cloisonnements et les obstacles qui freinaient la collaboration entre disciplines, entre organismes, entre recherche et industrie. Les conséquences de ce ratage ont longtemps pesé sur l'informatique française dans ses diverses composantes : enseignement supérieur, recherche, industrie.

L'informatique, fille des spéculations mathématiques et des besoins militaires, est née dans la recherche publique. Les premiers ordinateurs ont été construits, à la fin des années 1940, dans des universités et dans d'autres laboratoires d'État[1]. Cette expérience a ensuite été mise à profit dans l'industrie pour réaliser les premières séries commerciales. Le transfert technologique de la recherche à l'industrie s'est effectué de plusieurs manières. Parfois, il y eut adaptation du prototype "universitaire" par l'entreprise : par exemple les ordinateurs conçus à l'Université de Manchester ont été industrialisés par Ferranti. Souvent, des chercheurs sont passés dans le secteur privé, mettant leur expérience au service de nouveaux projets ; c'est ainsi qu'un ancien ingénieur du MIT ayant participé au projet SAGE, Ken Olsen, a fondé en 1957 Digital Equipment.

Ajoutons deux remarques à ce schéma descriptif. Les premiers ordinateurs ont été, dans leur grande majorité, construits au sein des départements de mathématiques, non des départements de "génie électrique" — significativement, à l'époque, on les appelle "machines mathématiques". D'autre part, ils n'y apparaissent pas *ex nihilo*, mais sur un terrain déjà préparé par l'existence, soit de calculateurs analogiques, soit de grandes machines électromécaniques, soit de services de calcul travaillant avec des machines de bureau : il est frappant de constater la corrélation entre la présence de ces équipements (parfois dès le début des années trente) et le développement ultérieur d'ordinateurs dans les mêmes institutions ou dans leur environnement immédiat.

1 Universités de Manchester et de Cambridge, National Physical Laboratory en Angleterre, universités de Princeton, de Pennsylvanie, MIT, National Bureau of Standards aux USA, etc.

Le même processus général se retrouve, avec un décalage de quelques années, en Europe continentale et au Japon. Or, de tous les pays industrialisés, la France est le seul où la recherche publique n'a réalisé aucun ordinateur à l'époque pionnière, c'est-à-dire avant 1960. Ce n'est pas faute que des projets aient été lancés. Comme nous le verrons, dans six centres de recherche français le processus décrit ci-dessus a été entamé, mais son déroulement s'est interrompu à un stade ou à un autre. On évoquera successivement ces échecs, leurs interprétations possibles, et leurs conséquences.

SIX ECHECS

L'Institut Henri-Poincaré

L'Institut Henri-Poincaré, créé en 1928, offre aux mathématiciens une organisation et des équipements collectifs (bibliothèque, services de calcul dotés de petites machines de bureau). En 1939, dans le contexte de la mobilisation scientifique, l'Armée charge l'IHP d'effectuer de nombreux calculs, notamment balistiques. Pour répondre à cette demande croissante (on sait qu'aux Etats-Unis le même besoin a conduit à la construction de l'ENIAC, le premier calculateur électronique, en 1945), deux laboratoires, l'un dirigé par Maurice Fréchet, l'autre par Louis Couffignal, conçoivent chacun une machine à calculer binaire, dont la construction est confiée à la société RBV. Le financement est assuré par les militaires, *via* le CNRS.

La débâcle de 1940 met fin à ces développements. Personnels et matériels se trouvent dispersés à travers la France ; le conseil d'administration de RBV, en majorité juif, est dissous ; les crédits de l'Armée et du CNRS sont réduits. Pourtant en 1942, Couffignal, devenu directeur du laboratoire de calcul mécanique du CNRS, négocie avec l'IHP : en raclant les fonds de tiroirs, on réunirait la somme nécessaire pour construire une seule machine, selon les plans de Couffignal, et moyennant une priorité donnée à Fréchet dans son utilisation. Les circonstances empêchent la réalisation de ce projet, mais le principe d'un monopole du CNRS sur la construction de calculateurs dans le milieu universitaire est désormais posé.

Le laboratoire de calcul analogique

Le premier organisme français qui réalise des calculateurs de quelque importance est l''Institut de mécanique des fluides de l'Université de Paris,

généreusement financé par l'Aéronautique. Dès 1932, le professeur Joseph Pérès et son assistant Lucien Malavard ont mis au point un procédé original de calcul, les "cuves rhéologiques" : dans une cuve pleine d'eau, parsemée d'électrodes, le courant électrique obéit aux équations de Laplace, les mêmes qui régissent certains phénomènes aérodynamiques. On peut ainsi déterminer le profil optimal d'une hélice ou d'une aile d'avion de manière plus rapide et plus économique que par des essais en soufflerie. Les "calculateurs Malavard" attirent des clients de plus en plus nombreux : les constructeurs d'avions français et britanniques, puis d'autres industries, notamment l'électrotechnique. Le laboratoire Pérès est en 1938 un centre de travail collectif, fréquenté par une quinzaine de chercheurs, ingénieurs, boursiers, thésards... Il bénéficie des moyens concentrés à la Cité de l'Air qui rassemble, place Balard, le Ministère de l'Air, l'Ecole Supérieure d'Aéronautique, divers "équipements lourds".

En juin 1940, l'équipe détruit ses équipements. Les travaux sont toutefois poursuivis pendant l'Occupation dans une semi-clandestinité, puis ils "redécollent" après la Libération. En 1945 le Laboratoire de calcul analogique est intégré au CNRS. Pérès devient directeur adjoint du CNRS, puis doyen de la faculté des sciences de Paris. Malavard est élu professeur à la Sorbonne. Tous deux siègent dans les hautes instances de la recherche et de l'enseignement supérieur.

Ce laboratoire a donc tout pour lui : des patrons prestigieux dominant une discipline qui a le vent en poupe, des "clients" dans l'Armée et la construction aéronautique, d'appréciables ressources financières et humaines. Comment expliquer qu'il ne construit pas d'ordinateurs, à l'époque où, à l'étranger, d'autres centres voués jusque-là au calcul analogique se "convertissent" au digital (le MIT par exemple) ? Trois raisons semblent constituer une explication suffisante :

- Pour Pérès et Malavard, le calcul n'est qu'un instrument, qu'ils ont fait progresser mais auquel ils ne consacrent pas l'essentiel de leur activité.

- Les cuves rhéologiques satisfont longtemps les besoins de la mécanique des fluides. Lorsque les performances des machines numériques surpasseront nettement celles des dispositifs analogiques, dans les années 1960, le laboratoire aura l'embarras du choix parmi les ordinateurs du commerce.

- En 1946, le CNRS décide de constituer un centre de mathématiques appliquées : l'Institut Blaise-Pascal. Celui-ci associe le laboratoire Pérès-Malavard, spécialisé dans le calcul analogique, et le laboratoire Couffignal, auquel est réservé le calcul numérique.

L'Institut Blaise-Pascal

Le but du CNRS, lorsqu'il fonde l'Institut Blaise-Pascal, est de développer les recherches sur les machines à calculer et de constituer un centre de calcul au service de la communauté scientifique. Création ambitieuse, puisque le CNRS voulait en faire «un centre de mathématiques appliquées unique au monde». A côté du laboratoire de calcul analogique, qui se développa normalement et existe toujours sous le nom de LIMSI, l'IBP comportait le Laboratoire de calcul numérique dirigé par Louis Couffignal, pionnier de la cybernétique.

Couffignal faisait figure, depuis les années trente, de spécialiste français de l'algèbre binaire et des machines à calculer, dont il ambitionnait de fonder la «science». En 1946, il est envoyé en mission aux États-Unis où il examine les premiers grands calculateurs, notamment l'ENIAC. Revenu en France, Couffignal se lance dans la réalisation d'une calculatrice électronique.

Ce projet baigne dans une étonnante ambiance de «chauvinisme technologique». "Nous avons au CNRS des spécialistes des machines à calculer, et il semble bien que leurs idées ont été utilisées par les Américains pendant la guerre" écrit le directeur du CNRS en 1946. "Grâce aux travaux de M. Couffignal, nous gardons une avance du point de vue théorique" déclare le Pr Pérès lors d'une conférence organisée par l'École supérieure des Télécommunications en juin 1947. Grande presse et publications savantes de l'époque célèbrent à l'envi «la machine française».

Ces cocoricos s'accompagnent d'un certain aveuglement quant aux investissements mis en oeuvre par les Anglo-saxons pour réaliser les premiers ordinateurs. Nommé inspecteur de l'enseignement technique en 1946, Couffignal ne peut consacrer, au mieux, qu'un ou deux jours par semaine à l'Institut Blaise-Pascal. Son laboratoire reçoit moins de 2 MF de crédits hors personnel sur les trois années 1947-1948-1949. L'effectif passe péniblement de 4 à 6 techniciens et un attaché de recherche. Couffignal mis à part, aucun membre du laboratoire ne va s'initier au calcul électronique en Angleterre ou aux États-Unis. Le CNRS ne s'est pas aperçu que sa "mise à niveau" dans ce domaine relevait avant tout d'un transfert de technologie et de savoir-faire. Et il ne s'en est pas donné les moyens.

De plus, les capacités professionnelles de L. Couffignal doivent être mises en question. Couffignal, qui n'est pas ingénieur, semble aussi ignorer les progrès de la logique formelle acquis pendant les années trente; et le concept de machine universelle contenu dans l'idée de calculateur à programme enregistré lui échappe complètement. L'architecture de Von Neumann, fondée sur un processeur arithmétique séquentiel et sur une mémoire aussi vaste que possible, lui paraît inélégante, coûteuse et typiquement américaine. La «machine française», au contraire, n'aura besoin de mémoire que pour les résultats intermédiaires des calculs (les instructions étant seulement gravées sur une pellicule lue par cellule photoélectrique), mais comprendra plusieurs processeurs fonctionnant en parallèle. Cette conception, au moins à l'époque, était une impasse technologique, et ne pouvait donner lieu à aucun développement de la programmation.

Le choix du partenaire industriel complète la liste des handicaps que le CNRS s'est donnés dans cette affaire. Le Centre confie la construction de la machine à la société Logabax, qui manque à la fois de compétence en électronique et de solidité financière. Une petite «machine-pilote» est bien présentée en 1951 (elle semble avoir effectué des calculs simples). Mais le

modèle définitif, à 2000 triodes, ne sera jamais construit, car Logabax dépose son bilan en 1952.

Couffignal ne fait donc pas du Laboratoire de calcul mécanique un centre de calcul équipé pour répondre à la demande des scientifiques français. En 1955, il se résout à acheter un ordinateur anglais qui servira peu, faute de personnel qualifié. Dès cette époque, des services de calcul concurrents fonctionnent chez IBM, Bull et au CEA, et la recherche se développe dans l'industrie et l'université : en dix ans, le CNRS a perdu son avance... théorique.

L'Institut d'Optique

En 1946, le petit bureau de calcul de l'Institut d'Optique ne suffit plus aux besoins de la recherche. Un jeune assistant, André Maréchal, réalise un intégrateur mécanique permettant de calculer des intégrations complexes. Et, comme c'est le cas à l'époque dans de nombreux laboratoires à travers le monde, cette machine ne fait que rendre les scientifiques plus conscients de leurs besoins en équipements de calcul performants. Cette situation est décrite dans le programme rédigé en 1950 par la direction de l'Institut d'Optique, qui envisage les "possibilités des machines à calculer rapides", c'est-à-dire électroniques. Le rapport stigmatise le retard de la recherche française en optique, gravement handicapée par la pénurie de moyens de calcul.

Maréchal élabore un projet de calculateur électronique et prend contact avec la Société d'électronique et d'automatisme, qui vient de se fonder précisément pour réaliser ce genre d'appareil. Des crédits et des postes sont demandés à cette fin au CNRS, avec le soutien de la commission d'Optique du CNRS. Mais celui-ci, accaparé semble-t-il par le projet de Couffignal (ou peut-être sous l'influence de celui-ci), ne donne pas suite. C'est finalement grâce à des fonds privés "et à l'appui de plusieurs grands services de l'État" que l'Institut d'Optique acquerra un petit calculateur IBM électromagnétique à cartes perforées, moins rapide qu'une machine électronique.

Le CNET

Le Centre National d'Études des Télécommunications est créé en 1944 avec de grandes ambitions mais de petits moyens. Organisme interministériel, il travaille à la fois pour les militaires (conception de radars, de crypteurs, de matériels radio, d'automatismes) et pour les PTT qui visent une reconquête par la recherche de l'indépendance nationale dans l'industrie des Télécommunications. Pourtant, jusqu'en 1954, le CNET est le parent pauvre de la science française. Ses crédits d'études diminuent alors que le prix des matériels subit une inflation dramatique. Plaie d'argent et entraves réglementaires l'empêchent de recruter et de retenir les personnels dont il a besoin sur le marché très tendu des électroniciens. Surtout, le CNET n'existe que sur le papier, il n'est qu'une fédération de laboratoires dispersés géographiquement, relevant de tutelles administratives distinctes.

Ces traits expliquent que le CNET ne construise aucun ordinateur avant la fin des années cinquante. Il en a pourtant les moyens techniques, et en éprouve la nécessité. Sur le plan technologique, il est sans doute le seul organisme en France à maîtriser toutes les compétences nécessaires à la réalisation d'un ordinateur : compteurs électroniques d'impulsions, lignes à retard, matériaux magnétiques, tubes spéciaux à haute fiabilité, mathématiques appliquées, etc. Par ailleurs, ses différents services ont un besoin croissant de calculs pour étudier les tubes électroniques et les antennes, comprendre le fonctionnement des réseaux, simuler le vol des missiles, mettre au point les filtres téléphoniques. Le "département de recherches mathématiques", longtemps réduit à un seul mathématicien (Louis Robin), est chargé d'élaborer ces problèmes. Mais, ne disposant pas de machines, il doit sous-traiter ses calculs à l'extérieur dans divers services spécialisés, tels que l'Institut Henri-Poincaré à Paris et la société Neyrpic à Grenoble. Les autres départements finissent, chacun de leur côté, par s'équiper de petits calculateurs et se tournent vers le CNRS pour demander des matériels plus puissants — en vain.

La situation se transforme en 1954. Le directeur du Service de recherche et du contrôle technique des PTT, Pierre Marzin, prend le pouvoir au CNET, en fait un véritable organisme unifié, impose une politique scientifique. Dès 1956 se dessine un axe de développement prioritaire : la commutation électronique, la commande par ordinateur des centraux téléphoniques. Mais, à cette date, les premiers ordinateurs commerciaux sont sur le marché (SEA, IBM, Bull) ; le CNET n'aura donc qu'un impact indirect sur l'industrie informatique française. En revanche, son expertise technique et sa stratégie avisée permettront à l'industrie française de la téléphonie, sous-développée et colonisée en 1950, de se hisser au premier rang mondial dans les années 1980. On retrouve ici le processus de transfert technologique de la recherche à l'industrie observé en informatique hors de France au cours des années cinquante.

L'ONERA

L'Office National d'Études et de Recherches Aéronautiques, fondé en 1946 avec d'importants moyens financiers, est l'un des premiers en France à investir dans les "équipements lourds" de calcul : appareils analogiques Pérès-Malavard, machines à cartes perforées Bull.

Cependant, de nouveaux besoins de traitement de l'information apparaissent avec les essais de missiles, qui exigent la reconstitution précise des trajectoires. L'ONERA entreprend en 1950 de mécaniser le dépouillement de ces télémesures, et réalise pour cela un ensemble d'appareils : le "Trajectographe".

Le besoin d'obtenir rapidement les résultats inspire l'idée d'y adjoindre un calculateur électronique. On désigne comme chef de projet un licencié en physique, qui étudiait des dispositifs haute fréquence pour l'aviation. Celui-ci se lance avec une petite équipe dans la construction d'une machine. Il s'agit d'un ordinateur extrêmement rapide pour l'époque (1 MHz). Trois

niveaux de mémoire sont prévus : une mémoire de travail instantanée (1 mot de 40 bits) constituée de bistables ; un tambour magnétique ; une mémoire de masse à bandes magnétiques. Le tambour est réalisé en enroulant autour d'un cylindre un fil de "Magnetophon" — procédé rustique mais relativement efficace. En revanche, la construction des dérouleurs de bandes magnétiques achoppe sur de multiples difficultés techniques. A part cela, en 1953 le calculateur semble en voie d'achèvement. Tout est remis en cause lorsque le chef de projet, en froid avec la nouvelle direction, part dans l'industrie.

Son successeur, un polytechnicien, fait en grande partie table rase du travail précédent et reprend à zéro un certain nombre d'études. Un apprentissage rigoureux est accompli en matière de fiabilité. Le nouveau calculateur, baptisé "MOP" (Machine Opératoire de la Physique), s'inspire de l'EDSAC construit à Cambridge. Ses 2 500 tubes et 6 000 diodes, occupant 16 armoires, effectuent une addition en 12 µs en virgule flottante : il est en principe aussi rapide que les plus puissants ordinateurs de l'époque. Mais l'ambition du projet et la faiblesse des moyens mis en œuvre par l'ONERA expliquent que la réalisation traîne. Les pires difficultés sont rencontrées dans la réalisation du nouveau tambour magnétique. Passons sur les avatars techniques pour en arriver à cet ahurissant résultat : au début de 1959, une "étape intermédiaire" (tambour de 512 mots) est atteinte, "permettant ainsi un fonctionnement d'ensemble de la machine MOP. Une nouvelle étape pourrait porter la capacité à 1024 adresses". A cette date, Bull, SEA et IBM construisent depuis trois ans, en France, des ordinateurs moyens à tambour de plusieurs milliers de mots.

En 1951, il était normal de concevoir un ordinateur dont la mémoire principale était un tambour ; dès 1955, les tores de ferrite s'imposent dans les grands calculateurs : progressant trop lentement, la "MOP" est dépassée techniquement avant même d'être achevée. Ses constructeurs en sont conscients et, en 1955, proposent soit de gagner du temps en achetant le tambour à une firme spécialisée, soit d'abandonner les circuits à tubes pour concevoir une machine à transistors et à tores de ferrites.

En 1957, IBM présente en France son gros ordinateur scientifique modèle 704 ; l'ONERA en est le premier acquéreur français, et l'achèvement du calculateur "maison" passe à l'arrière-plan. La MOP ne semble du reste jamais avoir été prioritaire dans les préoccupations de l'ONERA : l'Office dispose par ailleurs de moyens de calcul appréciables. Les télémesures dépouillées par le Trajectographe sont traitées sur des machines Bull dont la capacité centuple entre 1951 et 1957.

Le principal apport du travail accompli est une formation à la conception de circuits digitaux à tubes, donnée à des ingénieurs qui quittent l'Office pour diverses entreprises — le deuxième chef de projet est lui-même parti chez Bull en 1957 et a été remplacé par un troisième, qui démissionne à son tour en 1958 ! La machine finit par entrer en service, mais elle est nettement dépassée par les ordinateurs de l'industrie qui ne peut donc tirer aucun profit de cette expérience, même dans le domaine de la programmation.

TENTATIVE D'INTERPRETATION

Comment interpréter cet échec global de la recherche publique française à participer à la naissance de l'ordinateur ? Certes, chacun des ratages ci-dessus a ses causes particulières, contingentes. Mais on peut avancer trois explications générales : l'interruption du développement scientifique et technologique français par l'occupation nazie ; la situation des mathématiques appliquées en France ; les cloisonnements institutionnels et la faiblesse du *management* de la recherche dans notre pays, à l'époque.

La défaite et l'occupation

On retire de l'étude de la période 1938-1940 l'impression d'un rattrapage bien amorcé, brutalement interrompu par le *Blitzkrieg*. L'armistice de juin 1940 et l'occupation nazie ont une conséquence directe sur le développement des calculateurs en France, puisqu'ils obligent à abandonner plusieurs projets de machines à traiter l'information. Cela, non seulement à l'Institut Henri-Poincaré et au laboratoire de mécanique des fluides, mais aussi dans le milieu militaire où des projets de calculateurs de tir et de machines à décrypter les messages étaient en cours. La France est pendant quatre ans mise hors jeu, isolée du mouvement scientifique. La Deuxième Guerre mondiale, guerre des physiciens et des mathématiciens comme la première avait été celle des chimistes, se déroule pour l'essentiel sans le concours des savants français

De plus, la débâcle stoppe la mobilisation scientifique. Porteuse d'innovation dans la recherche, cette mobilisation avait stimulé la coopération interdisciplinaire, imposé le rapprochement entre recherche fondamentale et applications (mécaniciens des fluides et aéronautique, physiciens du magnétisme et Marine, etc.), orienté les universitaires vers la "science lourde" et suscité un début d'apprentissage de la gestion de projets technologiques.

La guerre fait aussi disparaître des hommes. D'une part nombre de savants français plus ou moins concernés par l'électronique ou le calcul sont tués par les nazis : Abraham, Bloch, Bruhat, Holweck et bien d'autres. D'autre part, si l'épuration fait peu de dégâts dans le milieu savant (les dirigeants du CNRS nommés à la Libération y veillent attentivement), les remous de la guerre amènent des scientifiques à quitter la France. C'est notamment le cas de Léon Brillouin, un physicien de tout premier ordre, spécialiste de physique des solides et de la théorie de l'information, qui en 1948 s'installe définitivement aux Etats-Unis, où il devient professeur à Harvard et conseiller scientifique d'IBM.

Enfin, la ruine du pays et le coût de la reconstruction mettent après 1945 les budgets de la recherche en concurrence avec des besoins vitaux à court terme et donc prioritaires.

Les mathématiques appliquées sous le boisseau

Cependant, la défaite de 1940 frappe un pays où la situation des mathématiques appliquées laissait déjà à désirer. Cette situation se retrouvera après la guerre et peut être caractérisée par deux traits : la domination des mathématiques pures, la faiblesse du calcul numérique.

Le "purisme" imprègne la science française du temps. Si l'utilité de la recherche pour le progrès de l'humanité est rituellement invoquée dans les discours académiques, dans la réalité on trouve peu de soucis d'applications. Cette conception est particulièrement forte en mathématiques, et s'accentue après la guerre sous l'influence du groupe "Bourbaki". Celui-ci, créé au milieu des années trente, domine les mathématiques françaises à partir de 1950 environ, quand ses membres fondateurs parviennent à des positions de pouvoir (chaires universitaires, commissions du CNRS et du CNU, Collège de France). Son but est d'établir les mathématiques sur des bases théoriques rigoureuses et de relever le niveau, assez médiocre entre les deux guerres, de l'école mathématique française. Ce brillant pôle d'attraction contribue à structurer le milieu mathématicien et lui impose une échelle de valeurs. Comme l'écrira Jean Dieudonné, il y a des mathématiques "nobles" et des mathématiques "serviles". Au sommet, l'on trouve des domaines comme la théorie des nombres ; à l'opposé, d'autres champs sont considérés comme indignes d'intérêt, c'est notamment le cas de l'analyse numérique. Déjà peu représentée en France avant l'offensive bourbakiste, l'analyse numérique fait ensuite figure de pis-aller, de voie de garage où un mathématicien ne peut faire carrière et être reconnu. Or l'analyse numérique joue un rôle capital dans le développement de l'informatique, qui à son tour va la renouveler profondément.

Sur un plan plus concret, le calcul numérique semble être lui aussi une activité sous-développée dans la France des années 1930-1955. L'inventaire de l'équipement des laboratoires en 1938 fait apparaître la pauvreté des moyens de calcul, qui freine la mobilisation technologique du pays à l'approche de la guerre. Quelques services de calcul existent — presqu'exclusivement à Paris —, dotés seulement de petites machines de bureau ; la seule exception est l'Institut de mécanique des fluides. Ces matériels seront du reste pillés par les Allemands. En compensation, après la guerre, la France impose à l'industrie allemande de livrer près d'un millier de "calculettes". Cette mesure permet de reconstituer rapidement le potentiel de recherche, mais à terme elle "bouche" le marché scientifique pour les fabricants nationaux.

Le problème nous paraît devoir être posé en termes de "demande", au sens économique du mot. Les chercheurs français ne semblent pas aussi "consommateurs" de calculs que leurs collègues anglo-saxons ou allemands. Après tout, si de puissants seigneurs de la science française d'après guerre, comme Néel ou Rocard, avaient eu besoin d'un ordinateur, ce n'est pas un Couffignal qui aurait pu les en empêcher. Or l'on n'observe rien de tel. Vers 1938 ils se contentaient de machines de bureau, alors que les principales universités d'Europe du Nord avaient déjà construit des répliques de l'intégrateur mécanique conçu par Vannevar Bush aux Etats-Unis. Dix ans plus tard, ils semblent se satisfaire d'équipements de calcul comparables à ce qui se trouvait avant la guerre à l'étranger (la seule exception, en France, est l'aéronautique, dont le soutien financier aux recherches sur les procédés de calcul est décisif).

Pourquoi nos scientifiques calculent-ils peu ? Suggérons trois interprétations, la première relevant de l'économie de la R&D, les deux autres de l'histoire des sciences. La recherche française a subi dans la décennie 1920 une "grande misère", ses crédits étant restés constants depuis 1914 tandis que l'inflation galopait ; la situation ne s'est rétablie que lentement au cours des années 1930 : les universitaires français ont été ainsi empêchés de prendre le tournant de la "science lourde", notamment d'entreprendre la construction de calculateurs. Par ailleurs, la physique telle qu'on la pratique en France à l'époque est essentiellement une physique expérimentale, réticente à adopter les modèles théoriques et mathématiques qui révolutionnent au même moment la conception de l'univers, à Copenhague, à Göttingen, à Cambridge, à Princeton[2]. Enfin, les scientifiques français disposent d'une excellente formation mathématique, qui leur permet de résoudre maints problèmes par des méthodes plus élégantes et moins fastidieuses que le calcul numérique.

L'un des exemples les plus probants à l'appui de notre hypothèse est la faiblesse de la demande de calcul au Commissariat à l'énergie atomique. Alors qu'aux États-Unis et en Angleterre la recherche atomique est un utilisateur pionnier des ordinateurs, en France le CEA se contente jusqu'en 1954 d'un modeste bureau de calcul, suffisant aux besoins immédiats du service de Physique mathématique. Ce bureau — quatre personnes utilisant des machines à calculer Friden — n'a à résoudre pour les physiciens et les ingénieurs du service que des pro-blèmes relativement simples, reflétant les modèles théo-riques de l'époque. "Ainsi dans le domaine de la neutro-nique, il était d'usage de se limiter pour certaines études en théorie de la diffusion, à des modèles à une variable d'espace (géométrie plane, sphérique ou cylindrique infinie), à deux ou trois milieux physiques distincts et à un ou deux groupes de neutrons. Les cas les plus complexes conduisaient rarement à plus d'une douzaine d'équations algébriques linéaires à une douzaine d'inconnues — ce qui était pourtant considéré comme un exercice redoutable par le bureau de calcul[3]."

En fait, dans la décennie qui suit la Libération, les seuls organismes qui ont explicitement d'importants besoins de calcul sont les laboratoires militaires. Ils les satisfont, soit en construisant eux-mêmes des calculateurs analogiques (généralement dérivés de machines allemandes), soit en finançant des

2 Pestre D. *Physique et physiciens en France 1918-1940,* Editions des Archives contemporaines, Paris 1984.

3 Amouyal A. "Les débuts de l'informatique au Commissariat à l'énergie atomique (1952-1972)" *Actes du 2e Colloque sur l'Histoire de l'Informatique en France*, CNAM, Paris, avril 1990.

développements dans l'industrie. Ils s'adressent moins facilement au milieu universitaire — nettement moins qu'en 1939, où l'urgence commandait. L'Armée ne stimule donc plus le développement du calcul numérique à l'Université et au CNRS, dans la décennie cruciale où sont construits les premiers ordinateurs.

Cloisonnements et blocages

Cela nous conduit à examiner les cloisonnements qui ont pu entraver la construction des "machines mathématiques". Depuis *La société bloquée* de Michel Crozier, une abondante littérature existe sur les défauts du "système français". Pour notre propos, il faut souligner les obstacles à la communication et à la collaboration entre diverses entités :

- Entre les disciplines scientifiques ; ainsi, il n'est pas question pour un laboratoire de mathématiques comme l'Institut Blaise-Pascal d'embaucher des électroniciens pour construire sur place la machine de Couffignal. Celle-ci est confiée à un sous-traitant extérieur, ce qui allonge les délais de réaction aux problèmes techniques (... mais permettra de décharger Couffignal de la responsabilité de l'échec). A l'étranger, les premiers ordinateurs sont construits au sein même des laboratoires de mathématiques : on n'y retrouve pas la réticence des mathématiciens français à fabriquer une machine. La question de l'organisation du CNRS (par disciplines, ou par thèmes ?) se pose à l'époque, et se résout en 1949 par l'adoption du cadre disciplinaire des universités. Le problème de leur interaction hantera le CNRS pendant des décennies.
- Entre l'Université (recherche fondamentale) et les Grandes Écoles (ingénierie), qui possèdent les capacités techniques mais ne font guère de recherche.
- Entre la recherche publique et l'industrie, coupure manifeste lorsque le CNRS choisit le partenaire industriel le moins capable de réaliser la machine de Couffignal, et lorsque l'ONERA monte son calculateur sans profiter de l'expérience acquise chez les premiers constructeurs d'ordinateurs français.
- En général, on constate le caractère très récent de la "science lourde" dans le monde scientifique français, et le manque d'expérience de la gestion de projets technologiques dans ce milieu, notamment chez les mathématiciens (ce manque apparaît nettement, par comparaison avec les laboratoires britanniques, américains et allemands). On ne peut que revenir sur le facteur le plus important, le plus fondamental à nos yeux : la défaite de 1940, qui a brutalement interrompu la "mobilisation scientifique" et l'apprentissage de la coopération entre savants, militaires et industriels, entamé depuis 1938.

Dans quelques universités de province (en particulier Grenoble et Toulouse), ces cloisonnements étaient peu marqués ; c'est précisément là que l'informatique s'est d'abord développée à la fin des années cinquante, mais en utilisant des machines achetées à l'industrie privée.

Dans les Armées, la volonté de développer et d'organiser les efforts de recherche en vue de rattraper le retard technologique français, attribué aux errements d'avant-guerre, est manifeste dès 1945. Mais elle se heurte à la fois au manque de moyens, à la difficulté de transformer les structures et les habitudes de travail, aux rivalités entre les corps et les armées, et sans doute à l'absence d'une véritable stratégie d'ensemble. Que l'on considère la coopération de la Marine avec Y. Rocard et L. Néel ou l'organisation des recherches dans l'Armée de l'Air, on ne trouve guère que de l'action au coup par coup. Dans les relations entre recherche et défense, la politique de contrats s'esquisse difficilement ; les "conseillers scientifiques" universitaires sont appelés par les militaires pour résoudre tel problème ponctuel, non pour participer à l'élaboration et au développement d'ensemble d'un projet. En 1945 est créé un Comité de coordination scientifique de Défense nationale, chargé de mettre en liaison les organismes de recherche civils et militaires. En fait, censé "coordonner" sans moyens financiers importants des structures qui restent inchangées au sein de chaque Armée, et sans véritable programme de recherche, ce comité est inefficient. Ce n'est qu'en 1962, avec la création de la DRME, qu'une véritable "coordination scientifique" pour la Défense nationale sera mise en œuvre.

CONSEQUENCES

Les conséquences de ces échecs ont été lourdes pour le développement de l'informatique en France.

Dans l'enseignement supérieur et la recherche, le manque d'ordinateurs jusqu'en 1955 a gravement retardé la formation de spécialistes; l'échec des projets de machines a sans doute retenu bien des personnes et des moyens matériels de s'investir dans l'informatique, domaine jugé dès lors à haut risque (peut-être doit-on faire remonter à cet épisode le faible développement de l'architecture d'ordinateurs en France); l'informatique n'a pas été prise au sérieux comme domaine de recherche autonome, et a été maintenue sous la tutelle des mathématiques jusqu'aux années 1970, avec des critères de sélection de plus en plus inadaptés ; enfin, cet échec n'a pu que perpétuer le fossé d'incompréhension ou de méfiance entre la recherche fondamentale et la recherche orientée vers l'ingénierie, et entre la recherche et l'industrie.

Faute de pouvoir bénéficier d'expériences en informatique acquises dans la recherche publique, les deux premiers constructeurs français d'ordinateurs, SEA et Bull, ont adopté des attitudes risquées :

- La Société d'électronique et d'automatisme, fondée en 1948, a effectué elle-même des recherches et des développements avancés, et assuré des activités d'enseignement, palliant la déficience du service public en ce domaine, mais au prix de son image d'industriel chez certains clients, ce qui freina son expansion.
- A l'inverse, Bull, ne trouvant dans le milieu académique ni réalisation technique éclairant les voies à suivre, ni spécialistes à recruter, suivit une stratégie trop conservatrice en ce qui concernait ses produits de grande diffusion (machines à cartes perforées), tout en laissant ses ingénieurs construire *in house* un chef-d'œuvre électronique mal adapté au marché, le Gamma

60. Il en résulta une grave crise, qui se termina en 1964 quand Bull passa sous le contrôle de General Electric.

Quant aux entreprises qui se sont lancées plus tard dans l'aventure informatique, elles ont emprunté une troisième voie, en produisant des ordinateurs de conception américaine. Ce choix handicapera le démarrage du Plan Calcul en 1967.

Pour prendre l'exemple précis des recherches sur la parole, deux remarques viennent à l'esprit. D'une part, l'apparition relativement tardive d'ordinateurs dans la recherche publique explique pourquoi la parole et l'informatique se sont rencontrées nettement plus tard qu'aux États-Unis ; on peut ici cerner précisément le phénomène du "retard français", trop souvent évoqué de manière floue. D'aure part, l'échec de la construction d'ordinateurs dans les six laboratoires étudiés a influencé la géographie actuelle des pôles de rencontre parole/informatique — sur le mode négatif : faute de machine, les recherches sur la parole ne se sont pas développées à l'IHP, à l'ONERA, à l'IBP... A quelque chose malheur est bon : ces six laboratoires étant tous situés en région parisienne, leur échec initial en informatique a favorisé le développement "décentralisé" de l'informatique française, et par là des recherches sur la parole.

Bibliographique indicative

Actes des trois colloques sur l'Histoire de l'Informatique en France (INPG Grenoble 1988, CNAM Paris 1990 et INRIA Sophia-Antipolis 1993).

Annals of the History of Computing , ACM.

Bashe, C. J. *et al.* 1986. *IBM's Early Computers*, MIT Press Series in the History of Computing.

Le Centre national d'études des télécommunications (1944-1974), F. du Castel & F. Lavallard éd., Centre de recherche sur la culture technique, Paris 1990.

History of Computing in the XXth Century, Actes du colloque de Los Alamos 1976, Academic Press 1980

History of Programming Languages, ACM, Academic Press, 1981.

Libois L. J. *Genèse et croissance des Télécommunications*, Masson, Paris 1983.

Pestre D. *Physique et physiciens en France 1918-1940*, Editions des Archives contemporaines, Paris 1984.

Vernay, J. *Chroniques de la Compagnie IBM France* IBM éd, Paris 1988.

Le présent article est essentiellement fondé sur le dépouillement des archives des organismes de recherche et des périodiques spécialisés.

TELECOMMUNICATIONS ET TRAITEMENT DE LA PAROLE : QUELQUES ELEMENTS HISTORIQUES

Michel CARTIER et Christian GAGNOULET
France Télécom - CNET
cartier@lannion.cnet.fr gagnoule@lannion.cnet.fr

Résumé

Le but de cette communication est de situer dans le temps quelques-unes des recherches sur la parole liées à la conception ou à l'exploitation des réseaux téléphoniques et d'esquisser un bilan des applications des technologies vocales dans les services de télécommunication.

On met en évidence les interactions entre les travaux sur la parole et l'usage qui en a été fait dans les réseaux téléphoniques. Jusqu'à maintenant, ces interactions ont surtout concerné les traitements au niveau acoustique, lesquels ont permis de produire des services. On pose pour le futur la question du même type d'interactions au niveau du traitement de la langue.

0. INTRODUCTION

La parole est la matière première des télécommunications : le poste téléphonique la saisit et la restitue, le réseau téléphonique la transporte en la transformant le moins possible. Les télécommunicants ont donc toujours porté un grand intérêt à la connaissance du signal vocal, de ses caractéristiques, de sa production, de sa perception. L'analyse-synthèse de la parole est une préoccupation fort ancienne, et les vocodeurs ont été dans une certaine mesure les précurseurs de la compression numérique.

Le moment semble opportun pour présenter quelques repères historiques communs au téléphone et au traitement de la parole avant d'esquisser un bilan de quelques applications des technologies vocales dans les services de télécommunication.

La première partie de cette communication montrera la richesse des interactions entre les recherches sur la parole, qu'elles soient à dominante fondamentale (production, perception, analyse-synthèse) ou tournées vers des applications (codage, synthèse, reconnaissance), et le monde des télécommunications. Dans la deuxième partie, en prenant appui sur quelques exemples récents, nous montrerons le parallélisme et les interactions entre les progrès en traitement de la parole, l'évolution des réseaux, l'émergence de nouveaux services vocaux et... l'existence des besoins. Pour terminer, nous comparerons le degré de maturité du traitement acoustique de la parole à celui de son traitement symbolique ou linguistique.

I. RECHERCHES DE BASE ET TELEPHONE

La première édition de l'ouvrage de Harvey Fletcher, "Speech and Hearing", est parue en 1929. Elle présentait le résultat de 15 années de travaux du Bell Telephone System, qui avait lancé dans les années 1910 un vaste programme de recherches sur la parole et sur l'audition. Des interactions entre les recherches sur la parole et la conception des réseaux téléphoniques se sont donc mises en place il y a fort longtemps. Si l'on remonte encore plus en amont, on remarquera que l'inventeur (ou l'un des inventeurs) du téléphone, Alexander Graham Bell, était phonéticien et spécialiste de la rééducation des sourds, à une époque où le moyen moderne de télécommunication était la télégraphie harmonique.

I.1. Alexander Graham Bell

En 1865, A.G. Bell était âgé de dix-huit ans. Il travaillait à la rééducation des sourds, comme son père, Melville Bell. Celui-ci avait découvert que certaines séquences de voyelles chuchotées donnaient l'impression d'une mélodie ascendante. Ces mêmes séquences évoquaient au contraire chez A.G. Bell une mélodie descendante. Il réalisa alors que "*each of the vowels was made up of two pitches, one rising in the sequence, the other falling*" (Bruce, 1973, p. 45). Il réussit à déterminer le "*pitch*" (hauteur globale) de certaines voyelles en plaçant des diapasons devant la bouche, expérience voisine de celles réalisées par Helmholtz pour déterminer les formants.

C'est ainsi, selon Bruce (op. cité, p. 46), qu'il fit les premiers pas vers l'invention du téléphone. Cette affirmation suppose que A.G. Bell avait en tête l'idée de transmettre non pas un signal microphonique considéré comme un tout, mais la superposition de signaux issus de plusieurs oscillateurs (op. cité, pp. 92-93). Cette idée résultait des deux éléments suivants : d'une part, les spécialistes du télégraphe s'intéressaient à l'époque au "*multiple telegraph*", le télégraphe harmonique, qui devait permettre de

transmettre simultanément sur le même fil plusieurs signaux de fréquences différentes dont chacun correspondait à un oscillateur modulé par tout ou rien ; d'autre part, Bell avait remarqué que les cordes d'un piano, comme les diapasons d'Helmholtz, pouvaient entrer en résonance avec certaines harmoniques de la voix. Ajoutons à cela qu'il avait découvert et expérimenté avec succès le fait que lorsqu'un sourd était capable de sentir au toucher la vibration des cordes vocales de son interlocuteur il pouvait aisément distinguer "p" de "b". Les éléments fondamentaux du Voder de Dudley étaient réunis, et l'on pourrait dire que le Traitement de la parole a précédé l'invention du téléphone.

I.2. Caractérisation et évaluation de la qualité

"*A comprehensive research program on speech and hearing, and its relation to the design of telephone systems*" (Fletcher, 1953). C'est dans ces termes que Harvey Fletcher présente les motivations des travaux du Bell Telephone System dans l'introduction d'un livre de quatre cent soixante pages qui résume les connaissances de l'époque sur l'audition et sur la production, l'analyse et la perception de la parole. On y trouve des données sur le spectre à long terme de la parole, sur la fréquence fondamentale et sur les formants ; des données sur les seuils d'audition et de masquage ; des méthodes subjectives et des méthodes objectives d'évaluation de la netteté. L'objectif d'ensemble était de contrôler les imperfections introduites par le réseau et de les conserver dans des limites acceptables.

Pierre Chavasse, fondateur en 1948 du Groupement des Acousticiens de Langue Française (GALF), l'actuelle Société Française d'Acoustique (SFA), est un pionnier de la science acoustique en France. Il est renommé pour ses travaux sur les postes téléphoniques et sur la téléphonométrie, commencés dans les années 20. Ses recherches sur la phonétique et sur l'intelligibilité (Chavasse, 1948) ont été poursuivis par René Lehmann (Lehmann, 1962) et par ses successeurs au département d'acoustique du CNET-Lannion (Combescure, 1981). On trouvera les dates-clés de sa carrière dans (Lorand, 1990).

L'évaluation de la qualité de la parole codée et de la parole synthétique figure depuis longtemps parmi les sujets auxquels s'intéressent les télécommunications, civiles ou militaires. En particulier, l'apparition des codeurs à 32 kbit/s a conduit plusieurs laboratoires à collaborer sous l'égide de l'IEEE dans le but d'aboutir à une procédure d'évaluation normalisée (Goodman & Nash, 1982). Ce travail a été relayé par le CCITT, organisme de normalisation de l'Union internationale des télécommunications (appelé maintenant UIT-T). Il a été l'amorce des recommandations P.81, P.82 et P.84, qui définissent des méthodes d'évaluation de la parole codée à divers débits et une dégradation de référence (le MNRU, Modulated Noise Reference Unit) permettant de calibrer les résultats. Les recherches se poursuivent pour trouver une référence mieux adaptée aux dégradations introduites par les systèmes récents de compression numérique. L'évaluation de la synthèse fait actuellement l'objet de recherches coordonnées dans le cadre d'instances diverses (Cocosda, projets européens SAM et Eagles). L'une des contributions des télécom à cette question est une recommandation du CCITT (UIT-T) qui traite la qualité des sorties vocales au niveau acoustique. Numérotée P.85, elle s'intègre dans la "série P.80" des recommandations concernant le codage mentionnées ci-dessus.

I.3. Analyse-synthèse, codage

Au cours des années 60 le vocodeur est apparu comme un moyen de transmission efficace. Est-ce dans cet esprit que Dudley a conçu son Voder ? Il semble que non si l'on se réfère aux titres de ses articles : "Synthesizing Speech" (Dudley, 1936), "PEDRO the VODER, a Machine That Talks" (Dudley, 1939a). Le Voder et le Vocoder étaient présentés comme des produits de la recherche et comme des outils au service de recherches ultérieures. On lit cependant dans l'article sur le Vocoder (Dudley, 1939b) que certaines propriétés des signaux pilotant le synthétiseur ouvraient peut-être la voie à une réduction de la bande passante nécessaire à la transmission de la parole...

Parallèlement aux premières études d'analyse-synthèse, les transmetteurs commençaient à envisager de numériser la parole. L'idée remonte au-delà de la première guerre mondiale : le premier brevet (français) date de 1938. Un système complet de transmission MIC à 8 voies est décrit par Black & Edson (1947) ; ce système comporte un codage à 5 bits avec un échelonnement logarithmique des niveaux. C'est finalement avec 8 bits par échantillon que le MIC actuel a été normalisé (CCITT, 1968).

En 1959, Gunnar Fant rappelait un travail qu'il avait effectué de 1945 à 1949 chez Ericsson, entreprise suédoise de téléphonie (Fant, 1959). Il écrivait : "*The acoustic theory of the speech process and the techniques of sampling and processing speech wave data have become of increasing importance for communication engineering science. The last few years' advance in methods of analysis-synthesis telephony are most promising in light of the very high compression rates that have been reached.*" G. Fant est connu pour ses recherches sur l'acoustique et la perception de la parole ; peut-être n'aurait-il pas écrit son fameux ouvrage "Acoustic Theory of Speech Production" s'il n'avait pas eu l'occasion de travailler pour une entreprise de téléphonie.

Les techniques de compression numérique n'ont pu progresser que grâce à une bonne connaissance des caractéristiques du signal vocal et de l'audition. Développer et discuter cette idée peut faire l'objet d'un long article, mais les bonnes synthèses sur le codage ne manquent pas (Calliope, 1989), et les repères sont connus (voir en bibliographie les recommandations du CCITT ainsi que Itakura & Saito, 1969 ; Atal & Schroeder, 1970 ; Esteban & al., 1978 ; Atal & Remde, 1982).

I.4. Commutation, informatique et parole

En 1962, le centre de recherches d'IBM à La Gaude entreprend des travaux sur les autocommutateurs téléphoniques. Il jouera un rôle important pour la communauté française de la parole : ses URV (unités à réponse vocale) et les outils d'analyse-synthèse associés ont fortement contribué au développement des recherches françaises sur la parole.

L'année 1962 voit également l'implantation à Lannion d'un centre de recherches du CNET dont le domaine d'activité essentiel, la commutation électronique, induit des recherches en informatique. De la rencontre entre l'informatique, la transmission numérique à très faible débit par vocodeur et l'acoustique téléphonique héritière de Chavasse naît la volonté d'entreprendre des recherches en traitement de la parole (reconnaissance et synthèse, puis codage), dont la première application doit être une URV rattachée au premier central de commutation temporelle "Platon".

Les premiers travaux de ces centres sont évoqués dans (Pertusier, 1993) et (Boë & al., 1993).

II. TRAITEMENT DE LA PAROLE ET SERVICES VOCAUX

Quatre-vingts ans après le début du programme décrit par Fletcher, on dispose de méthodes efficaces pour coder et comprimer le signal de parole téléphonique. Il n'y a donc plus de révolution à attendre sur le plan de la communication interpersonnelle directe.

La communication homme-machine sur support vocal, elle, n'en est qu'à ses débuts bien que des travaux sur ce thème se soient développés massivement dans les laboratoires dès les années 70 : les objectifs de recherche étaient ambitieux (parole continue, langue naturelle) mais ne visaient pas l'implantation rapide de services réels.

Au début des années 80, les premières tentatives d'exploitation de la commande vocale ont mis en évidence des difficultés insoupçonnées en laboratoire (importance des facteurs humains, de l'environnement acoustique, des exigences du grand public, etc.) et l'écart important qui existait entre les évaluations conduites en laboratoire et la réalité du terrain. Mais ces expérimentations, même infructueuses, ont été essentielles au développement des recherches en traitement de parole ; elles ont permis d'identifier les points bloquants et d'orienter efficacement les travaux ; elles sont sans doute à l'origine des progrès significatifs observés dans le domaine au cours des dix dernières années.

Les exemples suivants montrent l'émergence, grâce au traitement automatique de la parole, de nouveaux besoins et de nouveaux modes de communication dérivés de la communication interpersonnelle directe, mais aussi les multiples tâtonnements qui séparent l'idée d'un nouveau service de sa réussite commerciale éventuelle. Ils illustrent également la manière dont services et travaux de laboratoire se sont enrichis mutuellement en permanence au cours des années 80. Ces trois exemples, assez représentatifs, montrent le même cycle : des travaux de laboratoire préalables et l'apparition d'un besoin (réel ou supposé) suscitent des expérimentations de nouveaux services qui, à leur tour, font apparaître les insuffisances des techniques et alimentent de nouveaux travaux de recherche.

II.1. La messagerie vocale

La messagerie vocale est un exemple de service apparu pour répondre à un besoin de communication, et s'appuyant sur des technologies vocales parfaitement maîtrisées. Les évolutions des services reposent sur des progrès de la technologie des composants et des réseaux de télécommunication, plus que sur des progrès du traitement de la parole. La reconnaissance et la synthèse y restent aujourd'hui encore marginales, même dans le contexte des mobiles. Seules quelques expérimentations de la commande vocale sur le réseau fixe sont en cours ou prévues aujourd'hui (en Angleterre, en France, etc.).

L'idée de la boîte à lettres vocale est ancienne (Damlamian, 1974). Les progrès parallèles de la commutation, de la transmission et des techniques de compression de la parole ont fini par banaliser la communication interpersonnelle directe. La fonction de stockage numérique, liée aux techniques de compression, devenait vers 1975 disponible pour de nouveaux services. L'extension massive du service du téléphone a engendré un nouveau besoin : compenser l'indisponibilité de l'interlocuteur par une communication vocale différée. Ainsi est apparue la messagerie vocale, ou plus précisément le concept de service répondeur, qui n'en est qu'un des volets.

Ce besoin est apparu d'abord dans le monde professionnel, la messagerie vocale devenant rapidement indispensable aux entreprises. Le "répondeur" s'est installé en périphérie des commutateurs privés

Plutôt par un menu fixe
Un dialogue pour de clarification

(le plus souvent mal intégré, avec une ergonomie peu conviviale). Progressivement, des possibilités supplémentaires sont apparues (réception possible de plusieurs appels simultanés, accès distant aux messages à partir d'un terminal quelconque, etc.) et l'ergonomie s'est simplifiée.

Dans le grand public, l'introduction du service répondeur s'est faite au contraire au niveau du terminal téléphonique sous forme du classique "répondeur", analogique d'abord puis numérique. L'abaissement des coûts s'est accompagné d'une dégradation de la qualité de parole liée à des débits de codage de plus en plus faibles.

Aujourd'hui, le concept de messagerie passe par l'intégration dans les réseaux et apparaît comme intimement lié à leurs évolutions. Les services centralisés de messageries résidentielles sont appelés à remplacer les répondeurs en offrant à l'utilisateur des services plus conviviaux, plus riches et de meilleure qualité.

Le répondeur offre un service de communication interpersonnelle différée. La communication homme-machine n'apparaît que pour la commande du dialogue (réécoute des messages, enregistrement des annonces personnalisées, commande à distance, etc.), assurée par... le clavier téléphonique (disponible, bon marché et réputé fiable).

Un besoin spécifique de commande vocale est apparu pour les messageries dans les mobiles (difficulté, voire interdiction d'utiliser un clavier dans un véhicule). Le besoin a suscité quelques expérimentations qui ont immédiatement mis en évidence l'inadaptation quasi parfaite des systèmes de reconnaissance existants à ce contexte particulier (bruits, codages spécifiques). Aujourd'hui, de nombreux travaux de recherche visent à adapter la commande vocale aux exigences des mobiles. Il est à signaler que ces travaux sont menés surtout par des opérateurs et des industriels des télécommunications.

L'emploi de la synthèse a été proposé pour offrir une passerelle entre les messageries écrites et orales. Là encore, la confrontation avec la réalité a mis en évidence des difficultés techniques sous-estimées en laboratoire : des messages mal orthographiés, ponctués d'abréviations et de sigles, sont difficilement compréhensibles en parole de synthèse s'ils n'ont pas fait l'objet d'un prétraitement adéquat. Par contre, ces essais suggèrent un emploi réaliste de la synthèse pour la notification orale de messages écrits (transmission par synthèse de l'entête des messages), de même qu'elle est employée dans "Mirabel", système de diagnostic sur les lignes d'abonnés, pour transmettre vocalement aux opérateurs des informations relatives aux mesures effectuées.

II.2. Les services téléphoniques avec opérateurs

L'existence des opérateurs est associée de longue date au service téléphonique (réseau manuel). Les services par opérateurs restent très développés dans certains pays, aux Etats-Unis par exemple. Avec l'automatisation du réseau sont apparus des services complémentaires (renseignements, PCV, etc.) qui ont amené un changement du rôle des opérateurs et une augmentation de leur charge de travail.

L'utilisation des techniques de traitement de la parole dans ce contexte n'est pas réellement motivée par un besoin de l'usager (qui préférera encore longtemps un dialogue humain à un dialogue homme-machine), mais par une volonté d'optimiser les coûts d'exploitation. On peut ainsi décharger l'opérateur des tâches les plus répétitives et lui permettre de traiter un nombre accru de demandes.

En France, pour automatiser les services de renseignement (le 12), on a d'abord envisagé l'emploi de la synthèse. Mais la réduction du coût de stockage et la qualité alors insuffisamment naturelle des voix synthétiques ont finalement imposé l'usage de la parole codée. Le service d'assistance vocale aux opérateurs (AVO), aujourd'hui en cours de généralisation à France Télécom, permet à l'opérateur de transférer l'utilisateur vers un serveur dès qu'il a interprété sa question, le serveur se chargeant alors de vocaliser le numéro de l'abonné recherché et de conclure la communication. L'opérateur devient alors plus rapidement disponible pour traiter un nouvel appel.

Au Canada et aux Etats-Unis, cette volonté de réduire les coûts d'exploitation est allée plus loin, non sans impact sur les populations d'opérateurs. Depuis le milieu des années 80, plusieurs compagnies (AT&T, Northern Telecom, etc.) ont cherché à automatiser totalement certains services (le PCV par exemple) par l'emploi de serveurs vocaux (interaction par clavier ou par commande vocale). Aujourd'hui le grand public du continent nord-américain utilise régulièrement la reconnaissance de parole pour obtenir des services téléphoniques (plusieurs millions d'appels quotidiens).

L'orientation automatique d'appels est également possible grâce à la commande vocale. Cette fonction, répandue dans les entreprises (standards automatisés), est évaluée chez plusieurs opérateurs. En France, des expérimentations menées sur le 14 dans plusieurs agences commerciales ont amené une amélioration globale de la qualité de service.

Les enseignements issus de ces expérimentations ont conduit à des progrès sensibles de la commande vocale, en désignant clairement les points qu'il fallait améliorer prioritairement : rejet des entrées incorrec-

le poste de travail devient terminal téléphonique

tes, robustesse accrue, prise en compte des élocutions spontanées, etc.

Aujourd'hui, Northern Telecom cherche à aller encore plus loin en tentant d'automatiser davantage certains services de renseignements (utilisation de la reconnaissance de parole pour identifier certaines parties de la demande de l'utilisateur). Il ne s'agit pour l'instant que d'expérimentations ou de projets limités, mais il est probable que ces tentatives, quel qu'en soit le résultat, amèneront dans les années à venir de nets progrès des techniques de reconnaissance, comme les premiers essais sur l'automatisation des PCV l'ont fait vers 1990.

II.3. Répertoires et composeurs vocaux

Le répertoire à commande vocale est un exemple de service issu d'une possibilité technologique plus que d'un besoin réel (les répertoires existent depuis longtemps dans les terminaux). Cependant, l'histoire récente montre qu'un besoin réel nouveau peut apparaître indirectement à partir d'expérimentations plus ou moins réussies.

Une communication téléphonique directe est entièrement vocale ; pourquoi la phase de numérotation devrait-elle rester manuelle ? L'utilisateur, grâce à la commande vocale, peut établir une communication en prononçant simplement le nom de son correspondant. L'idée est simple, le contexte de reconnaissance l'est aussi (système monolocuteur, local, ambiance calme, etc.), et donc dès 1980 le problème est considéré comme potentiellement résolu.

Des expérimentations sont alors tentées. France Télécom fait développer deux prototypes de terminaux incluant un répertoire vocal dès 1980. L'un des deux atteindra en 1983 les rayons de la FNAC... où il restera en raison de son coût dissuasif. Le besoin n'est pas suffisant pour justifier le coût et les limites du produit apparaissent alors évidentes (ergonomie complexe, qualité de reconnaissance insuffisante, etc.) ; l'expérience française en reste là. L'idée n'est cependant pas abandonnée par tous. Des terminaux de faible coût apparaissent à la fin des années 80 (Audec, VCS, etc.), grâce aux progrès des composants. Aucun succès commercial important n'est cependant signalé.

La véritable nouveauté est venue des Etats-Unis en 1992. L'insertion du service au coeur du réseau change les données : coût quasiment nul pour l'utilisateur (abonnement), ergonomie fortement simplifiée (utilisation du téléphone et d'un serveur vocal centralisé pour programmer le système), nouvelles fonctionnalités (accès à partir de n'importe quel terminal), etc. En outre, la technologie, qui a bénéficié entre temps des expériences des services par opérateurs, est devenue beaucoup plus fiable. Les expérimentations menées par Nynex, US West ou Sprint semblent aujourd'hui des succès.

Un nouveau besoin est apparu avec ces expérimentations : la sécurisation de certaines communications rendue possible par l'authentification vocale (Sprint). Une nouvelle technologie vocale jusque là étudiée en laboratoire dans un contexte non téléphonique peut ainsi à son tour être confrontée à la réalité d'un service opérationnel. Là encore, un décalage est apparu entre les exigences d'un service opérationnel et la disponibilité des techniques. Il est encore trop tôt pour savoir si cette technique progressera suffisamment pour devenir réellement exploitable, mais il est certain que la recherche dans le domaine bénéficiera de cette confrontation avec le besoin du terrain.

III. CONCLUSION

Que conclure de ces quelques éléments historiques ? Les travaux fondamentaux lancés au début du siècle concernaient essentiellement la communication interpersonnelle directe (conversation téléphonique). L'évolution des réseaux téléphoniques, qui s'est traduite dans les années 60 par l'émergence du numérique en commutation et en transmission, associée aux travaux fondamentaux antérieurs sur l'analyse-synthèse de la parole, a permis de banaliser les techniques de traitement numérique de la parole. La généralisation du mode de communication interpersonnelle directe, couplée aux évolutions des réseaux de télécommunication, a suscité progressivement de nouveaux besoins et a provoqué l'apparition de nouveaux modes de communication : la communication interpersonnelle différée (liée à la technique de stockage dérivée de la compression), l'assistance à la communication directe grâce à des machines, embryon de communication orale homme-machine (techniques de reconnaissance et de synthèse de parole) et la communication mobile (techniques de compression, de traitement de signaux bruités et de reconnaissance de parole). Ces nouveaux modes de communication ont à leur tour permis le développement accéléré des technologies vocales, ce qui explique les progrès majeurs et décisifs au cours des dernières années, qui rendent enfin crédible l'usage de ces techniques dans des services réels.

Une première boucle est ainsi pratiquement bouclée, celle des traitements au niveau acoustique autorisant la communication interpersonnelle et, dans une ambiance calme, une forme élémentaire de communication homme-machine via la commande vocale. Cependant des travaux restent nécessaires pour lutter contre les perturbations acoustiques rencontrées en exploitation (bruit, écho, etc.). Les

connaissances fondamentales utiles concernant la production de la parole, sa perception et l'effet des dégradations de transmission sur la qualité du signal reconstitué sont en grande partie acquises, ainsi que les connaissances en traitement de signal et en traitements statistiques nécessaires à la conception d'algorithmes performants de compression numérique et de reconnaissance de parole.

Le traitement de la langue parlée naturelle (compréhension, synthèse à partir du texte, etc.) ouvre d'autres perspectives : celles d'une véritable communication orale homme-machine, bien sûr, mais aussi une nouvelle forme de communication interpersonnelle universelle quand les problèmes de traduction automatique auront été résolus. On en est encore loin. Mais certains projets actuels de recherche, ambitieux et à long terme, sont peut-être le point de départ d'une seconde boucle qui conduirait à la communication interpersonnelle multilingue. Rendez-vous en 2074.

BIBLIOGRAPHIE

ATAL B.S. & SCHROEDER M.R. (1970) : Adaptive Predictive Coding of Speech Signals, BSTJ, 49, pp. 1973-1986

ATAL B.S. & REMDE J.R. (1982) : A new model of LPC excitation for producing natural-sounding speech at low bit-rates, ICASSP, Paris, p. 614-617

BLACK H.S. & EDSON J.O. (1947) : Pulse Code Modulation, Telephony, 133-9, pp. 20-23 + 38

BOE L.J., GROSSETTI M., MOUNIER-KUHN P.E., CARTIER M., D'ALESSANDRO C., EMERARD F. & MELONI H. (1993) : Parole et Informatique en France : un quart de siècle après la rencontre, 3ème Colloque Histoire de l'informatique, Sophia-Antipolis

BRUCE R.V. (1973) : *Alexander Graham Bell and the conquest of solitude*, Cornell UP, New-York

CALLIOPE (1989) : *La parole et son traitement automatique,* éd. Masson et CNET-ENST

CCITT (1968) : Caractéristiques des systèmes MIC, Avis G.711

CCITT (1984) : Codage MIC différentiel adaptatif (MICDA) à 32 kbit/s, Recommandation G.721

CCITT (1988) : Codage du son à 7 kHz dans un canal à 64 kbit/s, Recommandation G.722

CCITT (1990) : Modulation par impulsions et codage différentiel adaptatif (MICDA) à 40, 32, 24 et 16 kbit/s, Recommandation G.726

CCITT (1992) : Codage de la parole à 16 kbit/s en utilisant la prédiction linéaire à faible délai avec excitation par code, Recommandation G.728

CHAVASSE P. (1948) : Essai sur la phonétique statistique de la langue française et son application à l'étude de l'intelligibilité d'une conversation, Annales des Télécom., Tome 3, n° 1, pp. 5-23

COMBESCURE P. (1981) : Vingt listes de dix phrases françaises phonétiquement équilibrées, Revue d'Acoustique, n° 56, pp. 24-38

DAMLAMIAN J.J. (1974) : Magnétophone analogique à étiquettes numériques visualisables, brevet français n° 2212072

DUDLEY H. (1936) : Synthesizing Speech, Bell Labs Record, XV-4, pp. 98-102

DUDLEY H. (1939a) : PEDRO the VODER, a Machine That Talks, Bell Labs Rec., XVII-6, pp. 170-171

DUDLEY H. (1939b) : The Vocoder, Bell Labs Record, XVIII-4, pp. 122-126

ESTEBAN D., GALAND C., MAUDUIT D. & MENEZ J. (1978) : 9.6-7.2 kbit/s Voice Excited Predictive Coder (VEPC), ICASSP, Tulsa, pp. 307-311

FANT G. (1959) : Acoustic Analysis and Synthesis of Speech with Applications to Swedish, Ericsson Technics, n° 1

FLETCHER H. (1953) : *Speech and Hearing in Communication*, D. Van Nostrand Company, Inc. (2ème éd.; 1ère éd.: *Speech and Hearing*, 1929)

GOODMAN D.J. & NASH R. (1982) : Subjective quality of the same speech transmission conditions in seven different countries, ICASSP, Paris, pp. 984-987

ITAKURA F. & SAITO S. (1969) : Speech analysis-synthesis based on the partial autocorrelation coefficient, Proc. JSA, pp. 199-200

LEHMANN R. (1962) : Etude psychophysique de l'intelligibilité du langage, Thèse, Fac. Sc. Paris

LORAND P. (1990) : L'acoustique, in *Le Centre National d'Etudes des Télécommunications,* Du Castel F. & Lavallard F., PUF-CRCT

PERTUSIER J.C. (1993) : Informatique et télécommunications ; 30 ans d'évolution à IBM, 3ème Colloque Histoire de l'informatique, Sophia-Antipolis

Remerciements. *- Ce travail a été effectué dans le cadre de la commission Parole et Informatique du GFCP de la SFA et du GTO Histoire, Parole et Informatique du PRC Informatique. Les auteurs remercient Louis-Jean Boë pour ses suggestions et pour sa riche et précieuse documentation ainsi que Christel Sorin et Pierre-Noël Favennec pour les améliorations qu'ils ont apportées.*

LE PREMIER CALCUL DES SERIES DE FOURIER

Jean-Bernard Robert
Centre de recherches
sur les très basses températures
CNRS
Avenue des Martyrs - 38042 Grenoble Cedex

Si pour le cristallographe, le spectroscopiste, ou l'ingénieur en mécanique l'analyse de Fourier fait depuis longtemps déjà partie de son langage quotidien, le chimiste, le médecin et le linguiste ne peuvent bien souvent pas se passer aujourd'hui d'en connaître quelques rudiments. C'est dans l'un des tous premiers travaux de la physique mathématiques, au début du XIX^{e} siècle, que sont apparus tout à la fois les notions de développement en série de Fourier et de transformées de Fourier, techniques à l'utilité universellement reconnue, à laquelle le développement des calculateurs rapides a donné un nouvel essor. Avant d'aborder l'exposé du problème physique qui a donné naissance à cette branche de l'analyse mathématique, nous présenterons à grands traits la vie de celui qui en fut l'initiateur.

Joseph Fourier fait partie de cette extraordinaire famille de mathématiciens et physiciens que connaît la France à la fin du XVIII^{e} siècle et au début du XIX^{e} siècle. Dans une liste forte incomplète et de ce fait injuste, nous pouvons citer Lagrange, Laplace, Monge, Haüy, Coulomb, Ampère, Galois, Sophie Germain... Fourier lui naît le 21 mars 1768, fils d'un modeste artisan tailleur. A dix ans il est orphelin, mais peut cependant par protection entrer au Collège militaire de sa ville natale où il se révèle un élève aussi doué pour les lettres que pour les sciences. Sa rhétorique terminée, il participe à l'enseignement du Collège, mais il lui faut faire un choix pour son avenir, c'est-à-dire dans sa situation se décider entre les Armes et l'Eglise. Il préfèrerait les Armes, ce lui lui permettrait de continuer à faire des mathématiques, mais il n'est pas noble, aussi rentre-t-il comme novice à l'Abbaye de Saint-Benoît-sur-Loire, où il va demeurer jusqu'en novembre 1789, dans un isolement scientifique qui lui pèse. Il commence cependant à l'Abbaye ses premiers travaux scientifiques, qui portent sur la résolution des équations algébriques. Les grands événements que connaît alors la France l'amènent à quitter Saint-Benoît-sur-Loire. En décembre, il présente à l'Académie ses premiers travaux.

Il regagne alors Auxerre où il est enseignant de lettres et de sciences au Collège. Jusqu'en 1793 il reste relativement indifférent au grand mouvement révolutionnaire, puis il sort de sa réserve au moment où la jeune République se trouve le plus menacée à l'intérieur comme à ses frontières. Il participe activement au comité de surveillance et au comité révolutionnaire d'Auxerre, chargé de délicates missions. Sa tendance modérée et son esprit de justice le fait sous la Terreur courir quelques dangers. Il connaît la prison et échappe de peu sans doute à l'échafaud, défendu par ses concitoyens d'Auxerre.

Fin 1794, lui que la fortune n'a jusqu'àalors que peu favorisé, va bénéficier de la création par la Convention de deux grandes institutions en matière d'enseignement, l'Ecole polytechnique et l'Ecole normale, qui allait plus tard devenir l'Ecole normale supérieure. En janvier 1795, désigné par le district voisin d'Auxerre de Saint-Florentin, Fourier se trouve à Paris, comme l'un des 1 400 élèves de l'Ecole de l'an III, où il va en sciences recevoir les leçons de maîtres prestigieux qui s'appellent Laplace, Lagrange, Monge, Haüy, Berthollet. Mais l'Ecole est un échec, l'écart est trop grand entre maîtres et élèves, cependant Fourier a été repéré par Monge et Berthollet en particulier, qui le font sans transition nommer enseignant à l'Ecole polytechnique en septembre 1795. Il va y enseigner la mécanique, des éléments de calcul différentiel et intégral, la géométrie descriptive. C'est l'époque où se prépare en grand secret l'expédition d'Egypte, qui aux 30 000 militaires doit s'adjoindre un fort contingent de scientifiques, Fourier est du nombre.

En juillet 1798, il débarque à Alexandrie. En Egypte, Fourier est administrateur, historien, archéologue et assure la lourde charge de secrétaire de l'Institut d'Egypte. Il ne lui reste que peu de temps pour faire des sciences. En novembre 1801, il rentre à Toulon «avec les derniers débris de l'armée française». Il aspire alors à une vie tranquille et toute entière consacrée à la science, sans encore cependant se définir comme littéraire ou

scientifique. Ses plans vont être contrecarrés, Napoléon le nomme préfet de l'Isère début 1802. Jusqu'en 1815 il va demeurer à Grenoble où il sera tout à la fois administrateur, historien et homme de sciences. Entre 1802 et 1812, il jette les bases de l'essentiel du message scientifique qu'il nous a laissé, sans rien négliger de l'administration de son département et en écrivant la Préface Historique du grand ouvrage que Napoléon a commencé sur l'Egypte.

Sans que l'on en connaisse les véritables raisons, Fourier aborde dès 1802 le problème de la propagation de la chaleur dans les solides. En 1807, il remet à l'Académie un premier manuscrit, puis un second en 1811 *«Théorie du mouvement de la chaleur dans les corps solides»*, qui porte en exergue une phrase en latin, qu'il nous dit venue de Platon et qui résume merveilleusement toute son oeuvre de physicien et de mathématicien *«Et ignem regunt numeri»*. Ce second manuscrit sera, pratiquement sans changement imprimé en 1822 sous le titre *«La Théorie Analytique de la Chaleur»*.

Le problème de physique auquel Fourier s'attaque en 1802 paraît simple : «lorsque la chaleur est inégalement distribuée entre les différents points d'une masse solide, elle tend à se mettre en équilibre, et passe lentement des parties les plus chauffées dans celles qui le sont moins ; en même temps, elle se dissipe par la surface, et se perd dans le milieu ou le vide. Cette tendance à une distribution uniforme, et cette émission spontanée qui s'opère à la surface des corps, changent continuellement la température des différents points. La question de la propagation de la chaleur consiste à déterminer quelle est la température de chaque point d'un corps à un instant donné, en supposant que les températures initiales sont connues». La simplicité de ce problème n'est qu'apparente. Il s'agit d'une question de physique qui recouvre de nombreux concepts à l'époque encore imprécis ou mal définis : chaleur, chaleur spécifique ; flux de chaleur, conductibilité en surface et en volume... Fourier va, héritier du philosophe Condillac, procéder d'une manière purement déductive, cherchant à traduire chacun des termes et concepts qu'il utilise dans la langue des calculs. Jamais il ne s'égare en images vagues ou imprécises des termes de physique qu'il utilise.

Pour la chaleur, dont il n'a pas pour son étude à approfondir la nature, il ne choisit pas son champ entre les tenants du calorique et ceux qui préfèrent la voir comme l'agitation des molécules d'un solide. Toute son oeuvre repose sur une seule hypothèse, celle de la transmission de chaleur de molécule à molécule. Si deux molécules ont pour température respectives v et v', et sont distantes de p, il passe pendant un temps dt de l'une à l'autre une quantité de chaleur $\Delta Q = (v'-v)\,\varphi(p)dt$. φ est une fonction dont il n'a nullement à définir l'expression analytique, elle est simplement décroissante avec p. La pierre angulaire de toute l'oeuvre de Fourier en physique, est l'introduction de la notion de flux de chaleur. Elle lui permet

Cette équation généralisée pour la propagation de la chaleur dans les trois directions dans un solide quelconque, lui permet d'écrire l'équation générale de propagation de la chaleur, universellement désignée sous le nom d'équation de Fourier :

$$F(\partial v;\partial t) = F(\kappa;CD)\ B(F(\partial^2 v;\partial x^2) + F(\partial^2 v;\partial y^2) + F(\partial^2 v;\partial z^2))$$

κ désigne la conductibilité interne, C la chaleur spécifique, D la masse volumique.

Cette équation établie, il faut la résoudre dans certains cas particuliers, abordée par Euler et Benouilli dans le cas des cordes vibrantes, cette résolution n'avait été qu'ébauchée. Fourier choisit un cas particulier simple, celui de la température stationnaire dans le solide limité par deux plans infinis à température nulle qui reposent sur une plaque à température 1. La distance des plaques à la température 0 est choisie égale à π.

En utilisant la méthode à l'époque bien connue de la factorisation de v en X(x). Y(y), il vient :

$$v = a\,e^{-x}\cos y + b\,e^{-3x}\cos 3y + c\,e^{-5x}\cos 5y + \ldots$$

où a, b, c... sont des coefficients inconnus à déterminer. Pour toute valeur de y avec x = 0, on a :

$$1 = a\cos y + b\cos 3y + c\cos 5y + \ldots$$

Fourier n'a toujours qu'une seule équation pour déterminer un nombre infini de coefficients. Il procède alors à des «différenciations successives», ce qui donne les résultats suivants :

$$1 = a\cos y + b\cos 3y + c\cos 5y + d\cos 7y + \text{etc...}$$

$$0 = a\sin y + 3b\sin 3y + 5c\sin 5y + 7d\sin 7y + \text{etc...}$$

$$0 = a\cos y + 3^2 b\cos 3y + 5^2\cos 5y + 7^2 d\sin 7y + \text{etc...}$$

La situation n'est guère confortable puisque «le nombre de ces équations est infini comme celui des indéterminées, a, b, c, d, e, ...etc...». Fournier choisit alors les sept dernières équations et au

terme d'un calcul par élimination mené avec élégance, qui consiste à multiplier la première équation par 11^2 et à lui soustraire la seconde, puis la seconde par 11^2 et à lui soustraire la troisième, etc..., il élimine g, puis f pour finalement obtenir :

$$a = F(3^2.5^2.7^2...;2.4^2.6^2...)$$

La symétrie de l'expression obtenue pour a lui permet de déduire ce qu'elle serait si, au lieu d'avoir choisi sept équations, il en avait pris un nombre infini. Puis faisant le rapprochement avec un résultat obtenu par Wallis, il trouve : $a = F(4;\pi)$

Au terme d'un calcul un peu long, où il montre son inclination pour le calcul algébrique, son organisation et son esthétique, il obtient tous les autres coefficients. Ainsi donc se trouve établie la relation qui présente le premier développement d'une fonction périodique en fonctions trigonométriques :

$$F(\pi;4) = \cos x - F(1;3)\cos 3x + F(1;5)\cos 5x - F(1;7)\cos 7x + F(1;9)\cos 9x + \ldots$$

Fourier, sans trop s'inquiéter de problèmes de convergence, examine le développement de nombreuses fonctions simples et écrit à propos de ses développements : «On peut étendre les mêmes conséquences à des fonctions quelconques, même à celles qui seraient discontinues et entièrement arbitraires». Ses démonstrations manquent de rigueur, ce qui lui sera reproché, par Laplace et Lagrange, qui ne sont pas eux-mêmes à l'abri de critiques. Mais l'heure en mathématique est plus à l'efficacité qu'à la rigueur.

Fourier applique ses résultats à de nombreuses fonctions périodiques, où il fait le calcul de ses coefficients, en utilisant non plus la méthode un peu longue adoptée dans son premier exemple, mais en adoptant la méthode aujourd'hui en usage. Puis passant à un système de longueur infinie, il introduit, sans en indiquer toute la signification, la transformée de Fourier.

Le reste de son oeuvre scientifique sera consacrée à l'application des méthodes qu'il a mises en évidence à de nombreux problèmes de chaleur.

LA MODÉLISATION ARTICULATOIRE : UN DEMI SIÈCLE D'ÉVOLUTION ENTRE FONCTIONNEL, PHYSIQUE ET BIOMÉCANIQUE

Louis-Jean BOË[1] Shinji MAEDA[2] Pascal PERRIER[1]

1. Institut de la Communication Parlée
CNRS URA 368, INPG/Univ. Stendhal, BP 25 38040 Grenoble cedex 9
(boe, perrier) @icp.grenet.fr

2. Laboratoire Traitement et Communication de l'Information
URA CNRS 820, Paris, ENST, 46 rue Barrault, Paris 13e
maeda@sig.enst.fr

Résumé

Domaine de prédilection de la phonétique au début du siècle, les études articulatoires ont acquis une nouvelle dimension scientifique grâce à la modélisation. D'abord fonctionnels, les modèles du conduit vocal tentent de plus en plus de cerner la réalité physiologique par des approches biomécaniques. Les modèles articulatoires offrent un cadre propice à l'inférence de stratégies de commande. L'étude de la modélisation du contrôle de la parole se déplace ainsi peu à peu du champ de connaissances du domaine traditionnel de l'acoustique et de la phonétique vers celui des sciences de la parole en relation avec le contrôle moteur et la robotique.

En France et au Japon les études articulatoires ont, nous l'avons relevé, une longue tradition, comme en témoignent travaux de recherche, réunions scientifiques et projets concertés.

Mots clés : *production de la parole, modélisation articulatoire, commandes, contrôle, histoire des sciences de la parole*

I. INTRODUCTION

C'est l'angle d'attaque articulatoire qui a permis aux recherches en parole d'acquérir, au début du siècle, le statut de science expérimentale. Le champ il est vrai se prêtait bien à une telle avancée : la phonétique pouvait alors récupérer les connaissances physiologiques mises à jour à la fin du XIX[e] siècle et les procédures d'observation et d'enregistrement que la médecine avait suscitées et développées (Marey, 1878 ; cf. l'ouvrage de F. Dagognet, 1987). C'est Pierre Rousselot qui reste, bien au-delà du paysage français, la figure emblématique de cette extraordinaire ré-appropriation dans le domaine de la phonétique expérimentale, comme il l'a exemplifiée dans ses *Principes* (1897-1908).

Les développements des études articulatoires ont ainsi précédé ceux de l'acoustique et de la perception qui sont montées en puissance avec l'émergence de l'électronique, de l'informatique et des méthodes de traitement de signal et d'analyse de données. Il serait d'ailleurs assez arbitraire de vouloir découper les recherches sur la parole en niveaux d'analyse articulatoire et acoustique, tant ils s'imbriquent et se renvoient, de façon récurrente, des interrogations fondamentales.

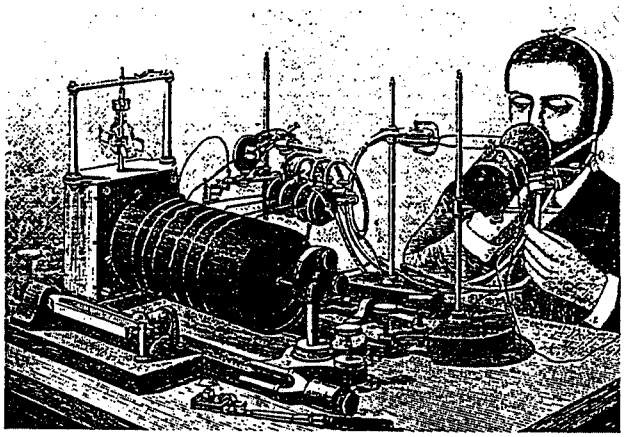


Figure 1. « 100 years ago : Abbot Rousselot's phonetic recorder ». In Scientific American, sept. 1992.

Le tour d'horizon des travaux articulatoires dans le domaine de la parole rassemblerait certainement plusieurs milliers de titres (remarquons au passage qu'il reste à faire), alors que l'intonation, la synthèse ou la reconnaissance de la parole... ont déjà fait l'objet d'un tel recensement. Pourtant, parmi ces recherches, la modélisation du conduit vocal ne représente, somme toute, qu'un nombre assez limité de publications. Ce n'est en rien paradoxal : la difficulté de regrouper toutes les données composantes, de mettre en œuvre un modèle, et d'en proposer une valorisation technique en sont en partie la cause. Nous serons quand même contraints de limiter ici notre revue des modèles dans le temps et dans le champ : nous n'avons retenu que les cinquante dernières années, nous avons laissé de côté les travaux sur le passage de la fonction d'aire à la fonction de transfert (qui relèvent de l'acoustique) et ceux de la simulation de la source vocale.

Tout en donnant une large part aux recherches françaises, nous nous proposons de faire une lecture des différentes approches de la modélisation, d'en souligner les multiples apports et de cerner les enjeux du futur.

II. DES LECTURES DE LA FONCTION D'AIRE

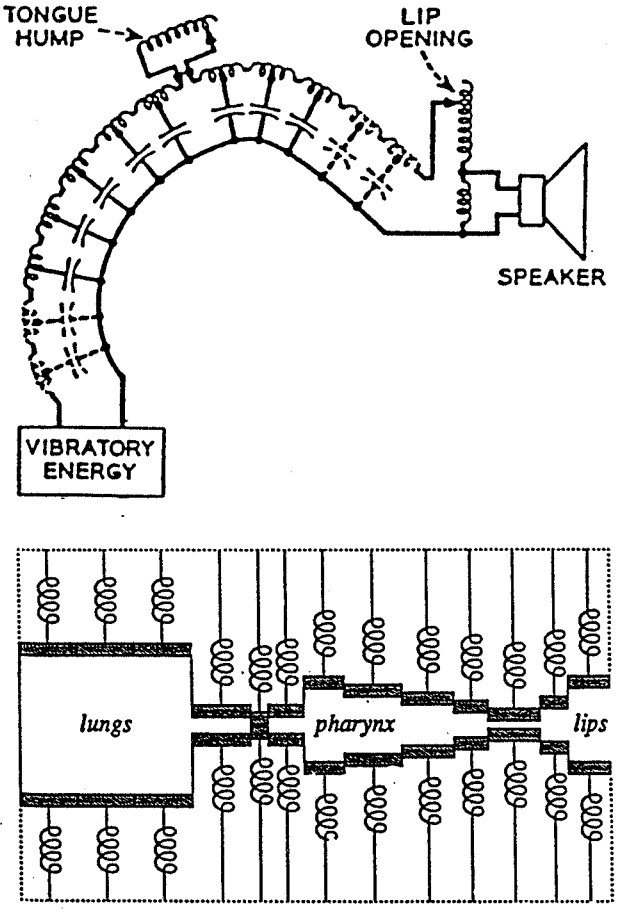
La caractérisation des sons du langage relève d'une longue tradition descriptive et classificatoire. C'est C.F. Hellwag (1781) qui est crédité du premier triangle articulatoire des voyelles et l'Association Internationale de Phonétique n'a procédé qu'à très peu de retouches depuis 1888, date de la première proposition, jusqu'à ces toutes dernières mises à jour. Pour ne nous en tenir qu'aux seules voyelles, trois traits articulatoires permettent de définir le quadrilatère centenaire : *avant/arrière* et *haut/bas* pour la masse de la langue, *arrondi/non arrondi* pour les lèvres. Ce qu'il faut retenir de cette proposition – dont il faut souligner la longévité et la résistance aux critiques et à certaines propositions divergentes venues de l'acoustique et de l'inversion – c'est essentiellement la simplicité du triplet des paramètres et l'unicité de la constriction intra-buccale. Il faut aussi noter que cette description 2D, définie par la position de langue, ne permet pas de passer à une simulation acoustique qui nécessite la quantification 3D du conduit aérifère. C'est en effet la fonction d'aire qui constitue l'entrée du passage à l'acoustique que les travaux de Chiba & Kajiyama (1941), de Dunn (1950) et de Fant (1960) ont permis d'asseoir théoriquement. Mais discrétiser le conduit vocal en n sections ne permet pas une commande économique : les 15 à 30 éléments nécessaires à une quantification satisfaisante présentant une trop grande complexité de commande et un manque de réalisme physiologique. On comprend assez bien que certaines modélisations articulatoires aient exploité au mieux les connaissances phonétiques séculaires et/ou les propriétés acoustiques de la simulation.

❒ *Intégrations de connaissances phonétiques*

Les travaux de Stevens & House (1955) et de Fant (1960, 1992) illustrent bien, à quelques variantes près, l'intégration des connaissances de la phonétique descriptive. Ces modélisations retiennent 3 paramètres : le lieu de l'unique constriction, sa dimension et l'aire aux lèvres. Cette approche récupère et prolonge une connaissance traditionnelle et ouvre tout un champ de nouvelles recherches : *if such specification is feasable it should be of theoretical and practical interest to students of speech in general. Linguists, phoneticians, speech therapists, etc., attempt to utilize articulatory to specify speech events* comme le souhaitaient Stevens & House. Nomogrammes caractérisant les relations articulatori-acoustiques, hypothèses théoriques de production vont en effet découler de ces modèles dont il faut souligner la simplicité et l'« intelligence » de l'approche.

❒ *Intégrations de connaissances acoustiques*

Dès le XVIII[e] siècle la simulation du conduit vocal sous forme de tuyaux a stimulé les acousticiens. La discrétisation, rendue nécessaire par la simulation du passage à l'acoustique, a renforcé ce type d'approche en accentuant l'aspect « tuyau » de la représentation. On trouve nettement cette tendance dans les présentations de Dudley (1955) de Barney & Dunn (1957), Leipp (1968) et Boersma (1993) Cette approche *n-tubes* va donner lieu à un certain nombre de travaux guidés par l'acoustique. Ceux de Bonder (1983a, 1983b, 1983c) vont permettre d'approfondir les relations fonction d'aire-formants et l'espace maximal. Les propositions de Mrayati, Carré & Guérin (1988), s'appuyant sur des propriétés de symétrie et d'antisymétries acoustiques de 8 tuyaux dérivées des fonctions de sensibilité différentielles du tuyau neutre, ont conduit les auteurs à proposer une nouvelle théorie globale de la production de la parole dont les limites articulatoires, acoustiques et phonétiques ont été relevées et discutées par C. Mokbel (1988) et Boë & Perrier (1990).



Figures 2. La conception « tuyau » : de Dudley (1955) à Boersma (1993).

Tout en notant l'intérêt acoustique évident de telles démarches il faut souligner la nécessaire relecture des fonctions d'aires qu'il ne faudrait manipuler qu'à la lumière des contraintes de production de la parole : elles ne peuvent être atteintes que par le déplacement des organes de l'articulation (et non l'inverse). C'est souvent faute de ce *caveat* que certains travaux sur l'inversion du conduit vocal ont obscurci la compréhension des relations entre espace articulatoire de commande et espace acoustique de sortie : si, théoriquement, il existe une multiplicité de fonctions d'aire équivalentes sur le plan acoustique, elles possèdent toutes (« modulo » les sensibilités locales) les mêmes zones de constriction.

III. DES ARTICULATEURS AU CONDUIT VOCAL

L'identité anatomique de certains articulateurs a favorisé l'approche modulaires : les lèvres, la mâchoire et les cavités nasales ont ainsi fait l'objet de travaux spécifiques, la langue restant la structure déterminante et centrale des modélisations globales du conduit vocal.

1. Les articulateurs

☐ *Les lèvres*

Extrémité visible et donc la plus accessible du conduit vocal, les lèvres ont été photographiées, chronophotographiées, cinématographiées et plus récemment filmées en vidéo classique et rapide (Lallouache, 1990, 1991 ; Rouy, 1992). La relation entre les dimensions intérolabiales (hauteur et séparation) et l'aire, proposée par Fromkin (1968) a toujours été confirmée. Les modèles 2D de face et de profil (Lindblom & Sundberg, 1972 et Abry & Boë, 1986, Guiard-Marigny, 1992) ont évolué vers des simulations 3D (Muller, 1975 ; Adjoudani, 1993) dans des contextes de synthèse de visages parlants (Benoît, 1992). Les modèles articulatoires permettent ainsi de tester au niveau des lèvres les conséquences perceptives des phénomènes de coarticulation et d'appronfondir les problèmes d'intégration bimodale (Robert-Ribes, Lallouache, Escudier, Schwartz, 1993).

☐ *La mâchoire*

Comme l'a bien souligné Bognar (1985), la mâchoire a longtemps été écartée des mesures et de la modélisation. Rares sont les phonéticiens du début du siècle qui ont fait preuve d'une réelle curiosité pour cet organe. L'abaissement-rotation mandibulaire provoquant une augmentation de l'aperture buccale, dans la plupart des cas reliée aux mouvements de la langue, c'est sur cette dernière qu'ont porté tous les efforts de description. Attitude renforcée par le fait qu'il est possible de parler (cas du fumeur de pipe) avec une mâchoire bloquée (confirmations effectives et plus scientifiques avec les expériences du *bite-block*). Corollairement, le rôle des déplacements antéro-postérieurs n'est pratiquement pas observé. Plusieurs modèles articulatoires se caractérisent par l'absence de paramètre mandibulaire (Öhman, 1964, 1966 ; Harshman, Ladefoged & Goldstein, 1977), ce qui a d'ailleurs soulevé une controverse (Mac Neilage, Ladefoged & Sawashima, 1979). Mais assez majoritairement les modèles articulatoires vont intégrer la mâchoire comme paramètre à part entière (Mermelstein, 1973 ; Perkel, 1974 ; Hiki & Oizumi, 1974 ; Fujisaki, 1977 ; Shiganaga & Arizumi, 1977) et étudier son influence spécifique (Lindblom & Sundberg , 1971 ; Maeda , 1978 ; Maeda, 1979 ; Majid, 1986 ; Majid, Abry, Boë, Perrier, 1987). La mâchoire est définie par la distance inter-incisive ; peut être choisi un centre de rotation approximativement au niveau du condyle (Mermelstein, 1973 ; Shirai & Honda, 1977) ou à l'arrière du canal dentaire (Hiki & Oizumi, 1974 ; Boë , Bognar, Gay & Descout, 1981). Les travaux de Gibbs & Masserman (1972), Bognar (1982), Edwards & Harris (1985), Edwards & Harris (1990) Bateson & Ostry (1992), et très récemment de Ostry & Munhall (à paraître) précisent les mouvements complexes de cet organe (rotation et translation) et se situent dans la problématique du contrôle. Véritable *pacemaker* de la parole, selon l'expression de Abry, la mâchoire joue peut-être même un rôle déterminant dans la durée universelle des syllabes (Brodda, 1979 p. 66 ; Lindblom, 1983 : autour de 160-170 ms, soit 6 Hz).

☐ *Le conduit nasal*

Il fait figure de grand délaissé de la modélisation. Les données anatomiques sont rares (Bjuggren & Fant, 1964), les moulages difficiles à réaliser vue la minceur des cavités, le passage à la fonction d'aire au niveau du voile du palais délicat à évaluer. Après quelques travaux préliminaires (House & Stevens, 1956 ; Hecker, 1962 ; Takeuchi, Kasuya & Kido, 1977) ce n'est que très récemment que les problèmes théoriques et leur simulation ont été réactivés (Maeda, 1982 ; Feng, 1986 ; Lonchamp, 1988 ; Castelli, 1989), mais, malgré quelques travaux (Moll & Daniloff, 1971), persite un manque évident de données dynamiques et l'intégration des fosses nasales dans un modèle de simulation globale reste à faire.

2. L'ensemble du conduit vocal

Pièce maîtresse du conduit vocal, la langue est une structure non rigide, très déformable (grâce à ses quelques 17 muscles intrinsèques et extrinsèques), dotée d'une grande vitesse de déplacement (en particulier son apex : Kuehn, 1973) fixée sur un plancher mobile (mâchoire et os hyoïde) et, à l'instar de la trompe de l'éléphant ou d'un tentacule de céphalopode, est entièrement libre à une de ses extrémités (Kier & Smith, 1985). Ajoutons à ceci qu'elle est difficilement observable par des techniques non-invasives et que les lectures radiographiques ne sont pas à la portée du premier lecteur venu. Décrite traditionnellement à partir de coupes sagittales, la langue a d'abord été modélisée dans son contour à partir d'analyse de Fourier, de traitement statistiques. Ce type d'approche, qui ne peut s'apprécier que par sa fonctionnalité n'a pas de prétention physiologique et il va falloir attendre le début des années 1970 pour voir apparaître des modélisations biomécaniques.

☐ *Modélisations fonctionnelles*

• Géométriques

Elles ont toutes en commun de simplifier le conduit vocal par une lecture géométrique de la coupe sagittale. Leur simplicité d'approche se heurte vite à des mauvais rendus physiologiques : la rigité circulaire de la langue ne peut prendre en compte sa capacité d'aplatissement-abaissement ; les mouvements consonantiques de l'apex nécessitant aussi une plus grande complexité. Mais cette même simplicité en a fait des instruments de recherche efficaces et développables à faible coût. Associés à un

synthétiseur à formants ils ont eu des visées pour la synthèse à partir du texte.

➳ Öhman (1964, 1966), Coker & Fijimura (1966), Coker (1967, 1968), Mermelstein (1969, 1973), Coker (1976), Nakajima, Omura, Tanaka & Ishizaki (1974), Ishizaki & Nakajima (1977), Ishizaki, Fuchi & Nakajima (1977), Heike (1977, 1980), Rubin, Baer & Mermelstein (1981), Hashimoto & Sasaki (1982), Wood (1982).

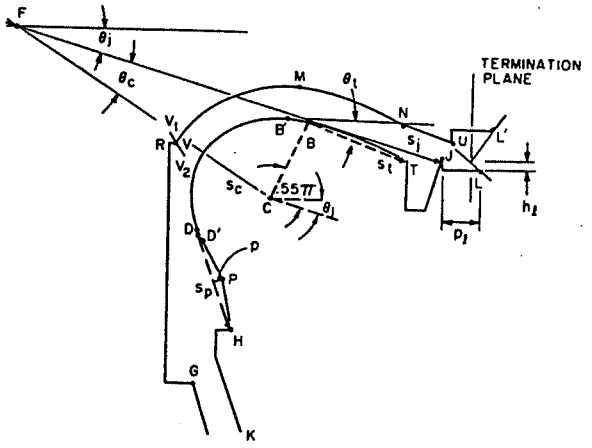


Figure 3. Un modèle très géométrique : Mermelstein (1973).

• Statistiques

Elaborés à partir de radiographies, les modèles ainsi regroupés permettent, par analyse factorielle, d'extraire les composantes principales de la coupe sagittale repérée sur une grille, et de décrire l'ensemble du conduit comme une somme pondérée de 5 à 7 vecteurs. Certaines de ces analyses sont guidées par une lecture physiologique qui ordonne l'extraction des composantes (méthode d'Overall par exemple). Plus proches de la géométrie effective du conduit vocal, ces modèles peuvent présenter les degrés de liberté des lèvres, de la langue (apex, déformation du corps et contraintes volumiques), de la mâchoire et de la position du larynx. Ces modèles, de la deuxième génération ont été exploités intensivement dans de multiples directions et souvent dans des contextes véritablement pluridisciplinaires.

➳ Lindblom & Sundberg (1971, 1974), Nakajima, Ohmura & Ishizaki (1975), Shirai & Honda (1976, 1977), Harshman, Ladefoged & Goldstein (1977), Maeda (1978, 1979, 1990), Gabioud (1994).

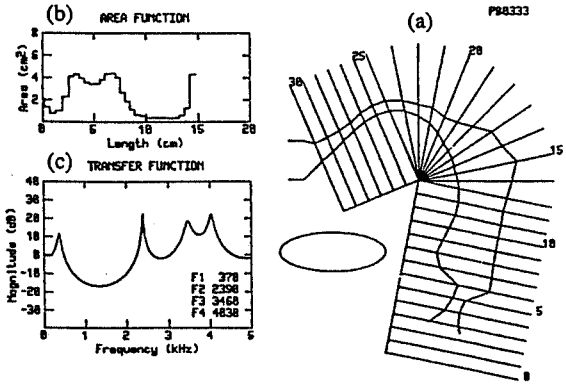


Figure 4. Un modèle statistique Maeda (1990).

❒ *Modélisations physiques et mathématiques*

Dans la veine de la formulation intuitive des modèles géométriques, Henke (1966, 1967a, 1967b) a élaboré un modèle descriptif de la forme de la langue qui décrit, en fait, les mouvement de points repères dans le plan sagittal médian (*flesh point*). Une caractéristique importante du modèle est que le déplacement de ces points est contraint par des lois physiques. En un sens on peut qualifier le modèle de Henke de modèle dynamique. Ces travaux n'ont pas eu de suite immédiate mais ils ont avancé un nombre important d'idées. Ainsi le modèle *look-ahead* continue à influencer les théories du contrôle, tout comme l'hypothèse du contrôle sélectif qui fait la différence entre le mouvement des points caractéristiques qui seraient directement contrôlés alors que les autres ne feraient que suivre ce mouvement.

D'autres modèles de la forme de la langue ont été élaborés avec une approche mathématique basée sur le calcul de paramètres à partir de données d'observation. Le premier de ces modèle est celui de Liljencrants (1971) qui propose une décomposition en série de Fourier. Les formes de la langue sont quantifiées à partir de coordonnées 1D dérivées d'une grille semi-polaire. Liljencrants met ainsi en évidence que les voyelles peuvent être décrite, avec une bonne précision, avec seulement les 2 ou 3 premiers coefficients de la décomposition de Fourier à partir de ces données monodimensionnelles. Les voyelles peuvent être ainsi représentées dans un « espace articulatoire » à deux ou trois dimensions. Assez étonnamment, ces résultats ont été confirmé, quelques années plus tard, par des méthodes d'analyses factorielles plus récente (par ex. Harshman, Ladefoged & Goldstein, 1977).

Avec les mêmes objectifs, Hashimoto & Sakai (1982) ont décrit les profils de la langue à l'aide d'une décomposition polynomiale quadratique. Ils sont ainsi caractérisés par cinq coefficients. Les auteurs ont proposé des relations polynomiales entre les formants et la forme de la langue correspondante et ils en ont dérivé des relations de correspondance (*mapping*) entre les espaces articulatoire et acoustique.

❒ *Modélisations biomécaniques*

Pour aussi intéressants qu'ils soient les modèles que nous venons de décrire, pas même celui de Henke, ne reflètent la structure biomécanique de la langue. Et pourtant il est tout naturel de penser que les caractéristiques physiologiques fondamentales de cette masse musculaire imposent des contraintes à ses déformations pour produire les sons du langage. La simulation biomécanique permet de mettre en évidence ces contraintes et la première tentative a été menée par Perkell (1974). Son modèle est construit à partir d'éléments musculaires visco-élastiques dont l'organisation reflète la structure musculaire de la langue décrite dans le plan sagittal médian. En fait ce type de simulation qui utilise le modèle musculaire-distribué (*lumped-muscle model*), avec les forces distribuées, préfigure les modèles à éléments finis.

Une des propriétés fondamentales de la langue qui ne prête pas à discussion, c'est bien celle de la conservation de son volume : la langue est pratiquement incompressible puisque les muscles contiennent un grand pourcentage d'eau. Mais seul un modèle 3D peut intégrer cette propriété fondamentale. C'est Kiritani, Miyawaki, Fujimura & Miller qui, en 1976, ont proposé la première simulation 3D à l'aide d'une méthode à éléments finis. Cette première tentative montre déjà que la contraction du génioglosse postérieur a pour effet de déplacer vers l'avant la masse de la langue tout en élevant sa partie antérieure. Cette déformation est due à deux contraintes : la conservation volumique de la langue et sa continuité élastique. La poursuite de cette recherche de modélisation 3D en éléments finis a conduit à deux séries de publications. Kakita & Fujimura (1977) et Fujimura & Kakita (1978) ont simulé l'ensemble du conduit vocal tout en proposant une version améliorée du modèle 3D de la langue, des lèvres et du larynx. Les fonctions d'aire et de transfert ont pu être ainsi calculées. Pour le [i], un effet de saturation acoustique a pu être mis en évidence en relation avec les forces musculaires et la déformation de la langue qui expliquant ainsi la stabilisation formantique de cette voyelle même si les forces de commandes augmentent. Cet effet de saturation se produit parce que la lame de la langue est poussée contre le palais dur. Ainsi une raideur suffisante de la lame peut permettre de maintenir inchangée la cavité en forme de tuile (*groove cavity*), ce qui est indispensable pour garder stable la production du [i] avec des efforts musculaires différents.

Par la suite, Kakita & Fujimura (1983), Kakita, Fujimura & Honda (1985) ont étudié, pour les voyelles, les relations entre les schémas de contraction musculaire observés par enregistrement EMG et dérivés du calcul du modèle commandé pour produire les formants correspondants. Les résultats montrent une bonne correspondance.

Hashimoto & Suga (1986) ont mis en œuvre un grand nombre d'éléments pour affiner la représentation du conduit (492 éléments au lieu des 86 des études précédentes) avec pour objectif d'approximer les coupes sagittales par une méthode des moindres carrés pour obtenir une estimation des schémas accentuels dans l'articulation de voyelles statiques à partir de la position neutre. En fait cette étude peut être considérée comme une démarche d'inversion : la récupération des tensions musculaires à partir des formes de la langue.

Ces modèles bio-mécaniques ont tous en communs d'être des modélisations statiques dans la mesure où ils ne font qu'intégrer l'élasticité, les contraintes dynamiques dues à l'inertie (ou la masse) ne pouvant être intégrées dans le modèle.

➳ Hiki, Niyada & Oizumi (1972), Hiki & Hoizumi (1974), Perkell, (1974), Kiritani, Miyawaki, Fujimura & Miller (1975), Kiritani, Miyawaki, Fujimura & Miller (1976), Fujimura (1977), Kakita, Fujimura & Honda (1985), Whilems-Tricario (à paraître).

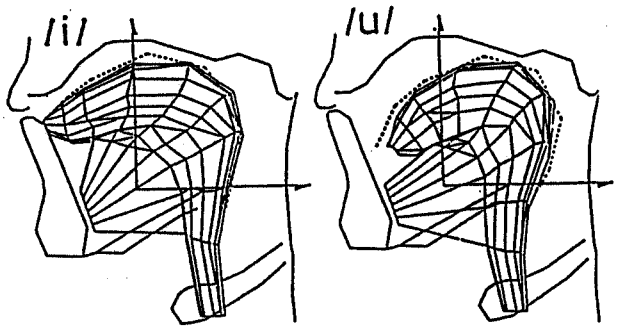


Figure 5. Un modèle à éléments finis Hashimoto & Suga (1986).

IV. APPORTS ET RETOMBÉES DE LA MODÉLISATION

La modélisation a focalisé et permis le développement d'un nombre fort substantiel de travaux qui relèvent de questions qui demeurent essentielles pour la compréhension du processus de production, de la mesure et de la représentation. Certaines applications pilote ont été avancées en synthèse et codage de la parole.

☐ *Affinements des techniques de mesure et de représentation*

L'énumération, par son ampleur, reflète la multiplicité des efforts qui ont été et sont déployés pour recueillir des données difficiles à extraire.

• Labiométrie et mandibulométrie

Les lèvres, parfois associées à la mâchoire ont suscité l'apparition d'une véritable technique : la labiométrie (photo, vidéo, radiographie, jauges de contraintes, capteurs opto-électroniques et magnétiques, lumière structurée, traitement d'image...) jusqu'à la mise œuvre de véritables environnements de traitement visage-parole. Il est vrai que les débouchés sont attractifs (synthèse du visage, dessins animés, transmission visiophonique, réalité virtuelle...).

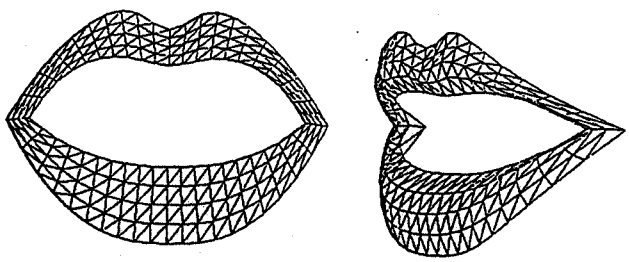


Figure 6. Un modèle 3D des lèvres Guiard-Marigny (1992) et Adjoudani (1993).

➳ Fromkin (1964), Ohala, Hiki, Hubler & Harshman (1968), Wioland (1971a, 1971b), Abbs & Gilbert (1973), Jankelson, Sawin, Crane & Radke (1975), McCutcheon, Fletcher & Hasegawa (1977), Muller & Abbs (1979), Louvion (1980), Abry, Boë, Corsi, Descout, Gentil, Graillot (1980), Sonoda & Wanishi (1982), Sonoda (1983), Gourret, Paillé, Boë & Descout (1985), Bothorel, Simon,, Wioland & Zerling (1986), Abry & Boë (1986), Lallouache & Worley (1988), Zerling (1990), Lallouache (1990, 1991), Rouy (1992), Maeda, Bothorel, Simon, Wioland & Zerling (1993), Shadle, Careter, Monks & Field (1993), Sonoda (1974).

• Représentation et mesure du conduit vocal

➳ **Moulages (ou dissection)** : Ladegoged, Anthony & Riley (1971), Bjuggren & Fant (1964), Sanchez & Boë (1984).

➳ **Radiographie**
Tomographie : Fant (1960).
Xérographie : Autesserre, Rossi, Sarrat, Giraud, Visquis, Demange & Chevrot (1979).
Radiocinégraphie : Simon (1967), Moll (1960), Perkell (1969).
Scanner : Perrier, Boë & Sock (1992).
Micro-Beam : Kiritani, Fujimura & Oshida (1971), Fujimura, Ishida H. & Kiritani S. (1968), Kiritani (1986), Abbs, Nadler & Fujimura (1988).

➳ **RMN** : Kiritani, Itoh & Fujimura (1975), Baer, Gore, Boyce, Nye (1987), Foldvick, Husby, Kværness, Nordli & Rinck (1991), Demolin & Segebarth (1992).

➳ **Ultrasons** : Minifie, Hixon, Kelsey & Wooshouse (1970), Sonies, Shawker, Hall, Gerber & Leighton (1981), Keller & Ostry (1983), Niimi & Shimada (1984), Honda (1985).

➳ **Electropalotographie** : Hiki & Itoh (1986), Marchal (1985).

➳ **Articulographie** : Höhne & al. (1987), Panagos & Strube (1987), Hoole & Gfoerer, (1990), Perkell & al. (1992).

➳ **Reconstruction dynamiques du conduit vocal en 3D** : Foldvik, Kristiansen & Kværness (1993).

➳ **Constitution de banque de données** : Bothorel, Simon, Wioland & Zerling (1986), Zohair (1993).

➳ **Appareillage et Postes de mesure** : de Kozhevnikov & al. (1964)... à Galindo & Teston (1989-1990).

Figure 7. Le « grand livre bleu » de Strasbourg (D'après Bothorel, Simon, Wioland & Zerling, 1986).

❒ *Avancées théoriques et des résultats*

• Pour la validation acoustique, la fonction d'aire est le passage obligé pour passer de la coupe sagittale à une estimation 3D. Les modèles articulatoires ont directement stimulé ces études (Heinz & Stevens, 1964 ; Sundberg, 1969 ; Ladefoged, Antony & Riley, 1971 ; Maeda, 1972 ; Perrier & Boë, 1989 ; Perrier, Boë & Sock, 1992 ; Beautemps, 1993). Une des premières retombées des modèles a permis tout naturellement de préciser les affiliations entre cavités, articulateurs et formants, de produire des nomogrammes, d'avancer la notion de focalisation.

➳ Fant (1965 ; Mrayati (1974), Fant (1975), Ladefoged & Harshman (1979), Ladefoged & Bladon (1982), Badin & Fant (1984), Boë & Abry (1986), Lin (1990), Badin, Perrier, Boë & Abry (1990), Boë , Perrier, Guérin & Schwartz (1990).

• Comme l'avaient prévu Stevens & House en 1955, a pu être mené bon nombre de descriptions phonétiques (nasalité, labialité, rétroflexion, constrictives, prototypes vocaliques). Il faut noter par exemple que la prototypisation des 37 voyelles de la basse de données phonologique d'UPSID n'aurait pas pu être avancée sans la modélisation.

➳ Lindblom (1972), Kakita & Fujimura (1977), Lindblom (1977), Nearey (1978), Hashimoto & Saski (1982), Goldstein (1983), Majid (1986), Majid, Abry, Boë & Perrier (1987), Jackson (1988), Boë & Vallée (1992), Kröger (1993).

• Analyse par la synthèse, synthèse articulatoire et perception, intégration visage-parole.

➳ Hafer (1974), Hiki & Harshman (1969), Werner & Haggard (1969), Purves, Blackett & Strong (1970), Hiki (1970), Shigenawa & Arizumi (1977), Rubin, Baer & Mermelstein (1981), Bailly & Schwartz (1991), Benoït, Lallouache, Mohamadi & Abry (1992), Mohamadi (1993), Le Goff (1993), Bothe, Rieger & Tackmann (1993), Robert-Ribes, Lallouache, Escudier & Schwartz (1993), Bothe, Rieger & Tackmann (1993), Gabioud (1994).

• La théorie quantique de Stevens (1972, 1989) complétée et largement discutée dans un récent numéro du *Journal of Phonetics* (n° 17, 1989) a été quantifiée et élaborée avec une modélisation du conduit vocal.

• L'inversion du conduit vocal et la récupération des gestes et le contrôle. Il s'agit en fait du problème central de contrôle de la production de la parole et des relations perception-action. Une longue mise à plat scientifique a permis de mettre en évidence qu'il existe une relation *many-to-one* entre répartition formantique donnée (formants et bande passante) et fonction d'aire. Sans contrainte, le problème de l'inversion est ainsi un problème mal posé au sens d'Hadamar. Et pourtant les sons du langage, pour un contexte donné, sont produit dans les mêmes zones du conduit vocal : il existe donc des contraintes de production qui permettent de retrouver par inversion ces solutions, d'où l'intérêt fondamental de disposer d'un bon modèle direct (le *plant*) pour pouvoir inverser. Ce problème est au cœur du projet ESPRIT

BRA *Speech Maps* (*Mapping of Action and perception in Speech*).

➳ Nakajima, Omura, Tanaka & Ishizaki (1974), Shirai & Honda (1976), Atal, Chang, Mathews & Tuckey (1978), Shirai & Honda (1978), Ladefoged, Harshman, Goldstein & Rice (1978), Hashimoto & Suga (1986), Shirai & Kobayashi (1986), Laboissière, Schwartz & Bailly (1990), Boë, Perrier & Bailly (1992), Bailly, Abry, Boë, Laboissière, Perrier & Schwartz (1993), Mac Gowan (1993), Kröger (1983).

☐ *Applications*

• Synthèse à partir du texte : précisons le bien, dans l'état actuel des connaissances, il ne serait pas raisonnable de prétendre rivaliser avec des synthétiseurs de type PSOLA par exemple. Pourtant dans les années 70 les laboratoires de la Bell ont mené plusieurs projets de synthèse à partir du texte (Flanagan, Coker, Rabiner, Schafer & Umeda, 1970 ; Coker, Umeda & Browman, 1973). Proposer un débit variable, passer facilement d'une voix d'homme à une voix de femme, la synthèse articulatoire permettra peut-être de produire de telles transformations auxquelles se prêtent mal les techniques de stockage de dictionnaires.

• Codage à bas débit : toujours à la Bell Telephone des études pilote de codage à débit réduit ont été avancées, avec des résultats de bonne qualité mais assez lourds à mettre en œuvre (Flanagan, Ishizaka & Shipley, 1980 ; Larar, Schroeter & Sondhi, 1988).

VI. QUEL CONTRÔLE POUR LES MODÈLES ARTICULATOIRES : VERS UN ROBOT PARLANT ?

La réédition de *Motor Phonetics* en 1988 a permis une véritable redécouverte du célèbre ouvrage, publié en 1951 par Raymond Herbert Stetson. La problématique posée par ce psychologue – véritable précurseur dans le domaine du contrôle de la production de la parole et très au fait des thèses de la phonologie et des travaux de la phonétique – reste une question fondamentale : comment sont structurées temporellement les commandes articulatoires pour produir la parole ? Les éditeurs, J. Kelso et K. Munhall, ont fait remarquer que si les lois Képlériennes permettent depuis le XVIII^{e} siècle d'expliquer le mouvement des corps célestes, il reste tout un champ d'interrogation pour interpréter et prédire les mouvements des corps terrestres.

Les modèles articulatoires offrent un cadre propice à l'inférence de stratégies réalistes de production de la parole, dont l'enjeu fondamental reste la compréhension des relations entre niveau symbolique et niveau physique. Ils constituent en effet un excellent outil d'évaluation quantitative d'hypothèses sur les coordinations interarticulateurs, sur la nature des variables contrôlées, en relation avec la commande linguistique. C'est en effet dans ce cadre que Henke a élaboré, dès 1964 un modèle articulatoire pour *improve our understanding of encoding of linguistic signals into acoustic signal through a study of the dynamic behavior of the vocal mechanism* (p. 8).

Une des théories les plus souvent adoptées en matière de contrôle et de prédiction de la coarticulation propose d'expliquer la variabilité contextuelle d'un son par la production en parallèle de différentes tâches articulatoires (Kozhevnikov & Chistovich, 1965 ; Öhman, 1966 ; Fowler, 1980). Cette théorie, formalisée par Fowler sous le nom de coproduction, a inspiré deux modèles célèbres de coarticulation : celui d'Öhman (1967) et le *Task Dynamic Model* des laboratoires Haskins (Kelso et al., 1986). Ce dernier a été mis en application en Europe, en particulier par Kröger (1993).

Öhman considère que l'articulation s'organise essentiellement autour des transitions entre voyelles (le canal de production vocalique) qui sont modifiées par la production des consonnes (le canal consonantique) selon un *timing* qu'il reste à définir. La variation temporelle de la forme du conduit vocal est ainsi essentiellement caractérisée par une transition continue d'une voyelle V_1 vers une voyelle V_2 avec effacement progressif de V_1 au profit de V_2 (*blending*). L'inertie des articulateurs est prise en compte par un filtrage du second ordre de commande-échelons. Öhman a pu ainsi retrouver et expliquer les trajectoires formantiques de parole naturelle dans des séquences V_1CV_2.

Avec la prise en compte des relations entre géométrie du conduit vocal, géométrie de la glotte et propriétés du signal acoustique s'affinent les propositions sur les objectifs de la tâche articulatoire : l'hypothèse de la coproduction est ainsi relancée dans un cadre élargi. C'est la perspective des laboratoires Haskins, qui autour de Saltzman, Kelso, Tuller, Browman et Goldstein (cf. par exemple Kelso et al., 1986) caractérisent l'activité articulatoire par rapport à quatre paramètres – les variables du conduit vocal : l'aire au lèvres, le degré de constriction palatale, le degré de constriction pharyngale, et l'aire glottique – qui définissent l'espace de la tâche (*Task Space*). La coarticulation est ainsi la conséquence de la production parallèle de gestes dans chacun des canaux définis par ces quatre paramètres. L'orchestration temporelle est alors régie par une *partition gestuelle* spécifiant les instants d'initiation et de fin des gestes (Saltzman & Munhall, 1989). Pour une *partition* donnée, la trajectoire finale dans l'espace des tâches est dépendante de la dynamique des articulateurs, qui est modélisée par un attracteur dynamique du second-ordre (concept de *Task Dynamic*). Le choix d'un modèle articulatoire commandé par des paramètres explicitement relatifs aux articulateurs (Mermelstein, 1973 ; Rubin et al., 1981) a permis d'associer, en amont de la partition, des séquences de commandes agissant directement sur les articulateurs. Ceci est rendu possible par la prise en compte implicite des relations entre articulateurs et variables du conduit vocal, et des coopérations possibles entre articulateurs pour la réalisation d'un objectif donné dans l'espace des tâches (*coordinative structures*). Kelso et al. ont ainsi pu rendre compte de certains phénomènes de compensation observés par Abbs & Gracco (1984) en réaction à des perturbations soudaines de l'articulation. Ce cadre modélisateur très puissant constitue ainsi l'outil fondamental d'évaluation et de mise à jour de la théorie de la *Phonologie Articulatoire* élaborée par

Browman & Goldstein (1986). La relation entre le niveau symbolique et le niveau physique est alors décrite avec le *geste phonétique*, comme combinaison de gestes élémentaires dans l'espace des tâches, comme base de la description phonologique.

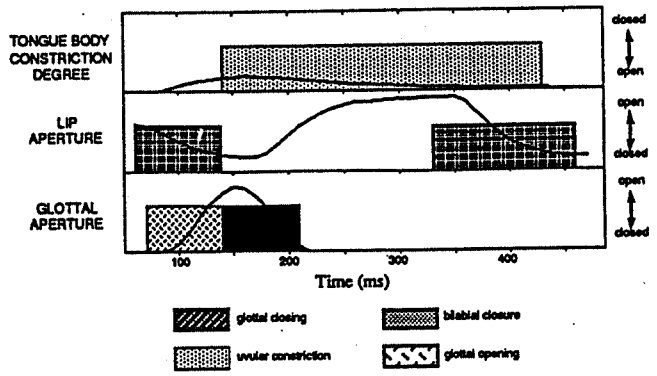


Figure 8. La partition gestuelle pour la synthèse de [pʌb] d'après Saltzman & Munhall (1989).

Le modèle des laboratoires Haskins, pour puissant qu'il soit, laisse cependant quelques questions importantes sans réponse : ainsi aucune proposition générale n'est avancée relative à la génération de la partition gestuelle, autre qu'un accord temporel ad hoc entre les différents canaux apte à rendre compte des signaux physiques mesurés (cf. sur ce point les perspectives de Saltzman & Munhall, 1989). D'autre part, différentes données articulatoires suggèrent l'existence de phénomènes d'anticipation gestuelle dépassant largement la limite du voisinage phonétique immédiat (Henke, 1966 ; Benguerel & Cowan, 1974 ; Perkell & Matthies, 1992 ; Lallouache & Abry, 1992 pour le geste de protrusion labiale). Il est ainsi vraisemblable qu'une partie des phénomènes de coarticulation soit le fait d'une planification spécifique au niveau cognitif. Dès 1966 Henke proposait d'en rendre compte par des règles de propagation de traits à tous les segments précédant le phonème labialisé, tant que la réalisation acoustique de ces segments n'en souffre pas. En fait, la diversité des observations faites par différents auteurs sur différents locuteurs (cf. en particulier Abry & Lallouache, 1993), ajoutée à la difficulté de structurer le continuum articulatoire et sonore en segments, plaident pour la recherche d'autres hypothèses. Une voie intéressante est proposée en 1983 par Lindblom puis affinée en 1990 : la production de la parole serait le résultat d'une négociation optimale entre un objectif de minimisation de l'effort d'articulation (optimisation de la tâche du locuteur) et celui de la réussite de la perception du message transmis (optimisation par rapport à la tâche de l'auditeur).

Ce concept a été exploité à l'ICP à Grenoble où un modèle d'inversion fondé sur des principes généraux d'optimisation a été élaboré (Bailly et al., 1991 ; Laboissière, 1992), en reprenant des concepts hérités du contrôle des gestes appris (Jordan, 1991 ; Kawato et al., 1991). Ici encore, les relations entre l'articulatoire et l'acoustique sont celles que propose un modèle articulatoire du conduit vocal (Maeda, 1979 ; Maeda, 1989), qui implicitement rend compte des contraintes articulatoires et des possibilités de compensation articulatori-acoustique. Des séquences de commandes articulatoires ont ainsi pu être inférées pour la génération de séquences vocaliques et de séquences du type VCV (Bailly & Laboissière, 1993), et les formes du conduit vocal associées ont montré de grandes similitudes avec des données radiographiques acquises sur des locuteurs. Par ailleurs, des stratégies de compensation à des perturbations stables de l'articulation (*bite-block* et tube labial) ont été établies, et les mouvements des articulateurs proposés correspondent bien aux données relevées sur des locuteurs placés dans des conditions similaires. Dans cette optique, le séquencement des mouvements articulatoires correspond à la spécification de cibles acoustiques en relation avec les différents phonèmes de la commande phonologique.

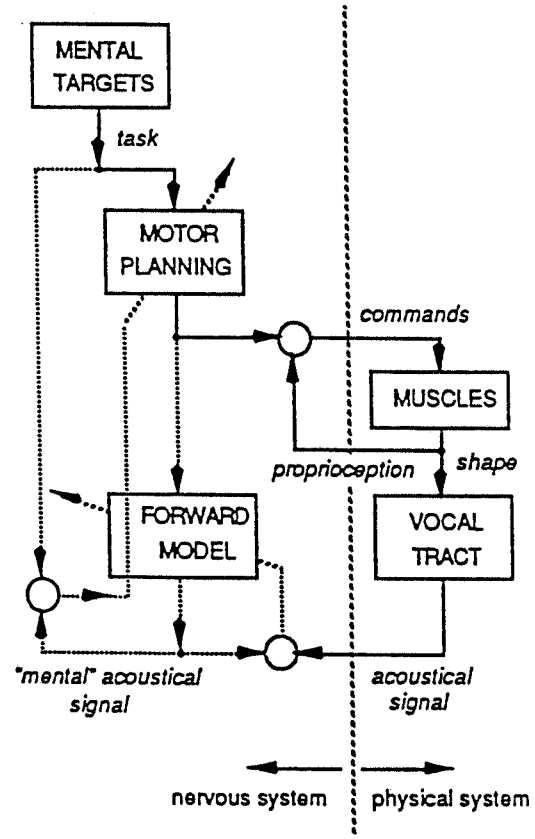


Figure 9. Une conception « robotique » de la parole (Laboissière, 1992).

Le passage des dispositions articulatoires (qu'elles soient inférées par l'une ou l'autre des méthodes) aux commandes motrices, est une nouvelle étape, abordée plus récemment, dont l'objectif est de se rapprocher toujours plus des niveaux supérieurs du contrôle de la parole, dans l'espoir de trouver une image toujours plus claire des commandes centrales associées à une chaîne symbolique donnée, et d'éclaircir le passage complexe d'une commande invariante à un support physique éminemment variable. Perkell (1974), Kakita et al.

(1985), Honda (1993), Maeda & Honda (1994) s'attachent à décrire les relations entre excitation EMG et dispositions articulatoires et/ou formants. Parallèlement, reprenant des concepts empruntés au contrôle des gestes appris (Théorie du *point d'équilibre* de Feldman, 1966), d'autres chercheurs considèrent les trajectoires articulatoires comme un cheminement entre cibles, séquentiellement ordonnées dans le temps et définies physiquement comme des positions d'équilibre mécanique entre les articulateurs (Flanagan et al., 1989 ; Loevenbrück & Perrier, 1993 ; Perrier & Ostry, 1994).

L'étude de la modélisation du contrôle de la parole se déplace ainsi peu à peu du champ de connaissances du domaine traditionnel de l'acoustique et de la phonétique vers ceux du contrôle moteur et de la robotique. En effet la similitude des problèmes rencontrés en modélisation du contrôle de la parole avec ceux auxquels sont confrontés les roboticiens va croissante : nécessité de réaliser un *plant* rendant compte de la réalité des relations physiques entre le robot et le monde physique (le modèle articulatoire du conduit vocal, les modèles d'articulateurs), questions fondamentales sur la représentation du monde physique en termes interprétables par le robot (définition de l'espace de la tâche), caractérisation des relations entre action et monde physique (relation action-perception), optimisation de la tâche.

VII. SANS VOULOIR TIRER LA COUVERTURE ARTICULATOIRE... FRANÇAISE ET JAPONAISE

En France les études articulatoires ont, nous l'avons relevé, une longue tradition : l'abbé Pierre Rousselot a suscité en Aix-en-Provence, Grenoble, Lannion, Strasbourg, de nombreux émules, voire des vocations... Témoins de cette activité, plusieurs colloques internationaux ont été organisés au cours de ces vingt dernières années dans le cadre de la communauté parole et plus particulièrement du Groupe Francophone de la Communication Parlée. Citons pour mémoire en 1977 le séminaire *Articulatory Modelling and Phonetics* (Grenoble : R. Carré, R. Descout & M. Wajskop), en 1979 le séminaire *Larynx & Parole* (Grenoble : L.-J. Boë, R. Descout & B. Guérin), en 1980 l'*International Seminar on Labiality* (Lannion : R. Descout, R. Besson, L.-J. Boë, C. Abry) en 1988 le *1rst Seminar on Speech Production Models and Data* (Grenoble : C. Abry, L.-J. Boë, B. Guérin) et en 1989 le *Speech Production and Speech Modelling. Nato Advanced Study Institute* (Bonas : W.J. Hardcastle & A. Marchal). Au cours des années 1980 le département RCP du CNET, sous la responsabilité de R. Descout, a joué un rôle déterminant : par ses études internes et par sa politique de convention avec les Instituts de Phonétique d'Aix en Provence, Grenoble et Strasbourg et par son animation au sein du Groupe Francophone de la Communication Parlée (campagne de recueil de données : labiofilms, radiocinégraphie, signaux EMG, signaux kinésiographiques ; traitement des données et modélisation). Avec le développement des contrats européens, il n'est donc pas étonnant que l'on retrouve deux laboratoires français à la direction de projet ESPRIT *BasicResearch : Articulatory-Acoustic Correlations in Coarticulatory Processes : A Cross-Language Investigation* ACCOR (9 laboratoires, *prime* l'Institut de Phonétique d'Aix, A. Marchal) et *Speech Maps : Mapping of Action and Perception in Speech, Sound-to-Gesture Inversion in Speech* (14 laboratoires, *prime* l'Institut de la Communication Parlée, C. Abry & P. Badin).

Il serait *scientically incorrect* de ne pas souligner la très prolifique « filière japonaise » qui enrichit la recherche scientifique de travaux à la fois dans le domaine de la mesure, de la physiologie et de la modélisation (cf. la bibliographie) : l'ère moderne de la modélisation n'a-t-elle pas été ouverte, en 1941, par les travaux d'un phonéticien et d'un électronicien japonais Chiba & Kajiyama ? Est-ce un hasard aussi si l'un des auteurs est le plus français de tous les japonais travaillant dans ce domaine !

BIBLIOGRAPHIE

Abbs J.H. & Gilbert B.N. (1973) A Strain Gauge Transduction System for Lip and Jaw Motion in Two Dimensions : Design Criteria and Calibration Data. J. of Speech and Hearing Research, 16, 248-256.

Abbs J.H. & Gracco V. (1984) Control of Complex Motor Gestures : Orofacial Muscle Response to Load Perturbation of Lips during Speech. J. of Neurophysiology, 51, 705-723.

Abbs J.H., Nadler R.D. & Fujimura O. (1988) X-Ray Microbeams Track the Shape of Speech. SOMA, 29-34.

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

Abry C., Boë L.-J. & Schwartz J.-L. (1989) Plateaus, Catastrophes and the Structuring of Vowel Systems. J. of Phonetics 17, 47-54.

Abry C. & Feng G. (1991) Peut-on entendre les formes du conduit vocal ? Le Courrier du CNRS. Dossiers Scientiques, Signaux et Images, 95-96.

Adjoudani A. (1993) Élaboration d'un modèle de lèvres 3D pour animation en temps réel. DEA Signal Image Parole, INP Grenoble.

Al Ansari A. (1981) Étude du fonctionnement et simulation en temps réel d'un modèle de source vocale. Étude des interactions source-cavité subglottiques et source-conduit vocal. Thèse Docteur Ingénieur. Inst. National Polytechnique, Grenoble.

Alfonso P.E.K., Honda K. & Baer T. & Harris K.S. (1982) Multichannel study of tongue EMG during Vowel Production. J. Acoust. Soc. Am., 71 S54(A).

Angola O. (1992) Analyse labiométrique en temps réel par traitement d'images vidéo. DEA Signal Image Parole, INP, Grenoble

Atal B.S., Chang J.J., Mathews M.V. & Tuckey J.W. (1978) Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer Sorting Technique. J. Acoust. Soc. Am. 63, 1535-1555.

Autesserre D., Rossi M., Sarrat P., Giraud G., Visquis R., Demange R. & Chevrot L. (1979) Exploration radiologique de l'oropharynx et du larynx en phonation. Séminaire Larynx & Parole, 47-73.

Badin P. & Boë L.-J. (1987) Vocal Tract Nomograms : Acoustic Considerations. A Crucial Problem : Formant Convergence. 12th Int. Congr. Phonetic Sciences, 2, 352-355.

Badin P. & Fant G. (1984) Notes on Vocal-Tract Computation. STL QPSR 2/3, 53-108.
Badin P., Perrier P., Boë L.-J. & Abry C. (1990) Vocalic Nomograms : Acoustic and Articulatory Considerations upon Formant Convergences. J. Acoust. Soc. Am., 78, 1290-1300.
Baer T., Alfonso P.J. & Honda (1988) Electromyography of the Tongue Muscles during Vowels in /əpV p/ Environment. Annual Bulletin of RILP (Univ. Tokyo), 22, 7-19.
Baer T., Gore J.C., Boyce S. & Nye P.W. (1987) Application of MRI to the Analysis of Speech Production. Magnetic Resonance Imaging, 5, 1-7.
Baer T., Gore J.C., Gracco L.C. & Nye P.W. (1991) Analysis of Vocal Tract Shape and Dimensions using Magnetic Resonance Imaging : Vowels. J. Acoust. Soc. Am., 90, 799-828.
Bailly G., Boë L.-J., Laboissière R., Perrier P. & Schwartz J.-L. (1993) Inversion in Speech Recognition. In *Signal Processing VI : Theory and Applications*, J. Vandewalle, R. Boîte, M. Moonen, A. Oosterlinck ed., Elsevier, Amsterdam.
Bailly G. & Laboissière R. (1993) Learning coarticulation and compensation for selected VV with a control model. In *First Annual Report - WP3 Dynamic constraints and motor controls*, G. Bailly ed., Speech Maps Project (ESPRIT BR Project n° 6975).
Bailly G., Laboissière R. & Schwartz J.-L. (1991) Formant Trajectories as Audible G esture : An Alternative for Speech Synthesis. J. of Phonetics, 19, 9-23.
Bateson E.V. & Ostry D.J. (1992) Rigid Body Reconstruction of Jaw Motion in Speech. J. Acoust. Soc. Am., 92, 2391.
Beautemps D. (1993) Récupération des gestes de la parole à partir des trajectoires formantiques : identification de cibles vocaliques non-atteintes et modèles pour les profils sagittaux des consonnes fricatives. Doctorat INP Grenoble.
Beautemps D. & Gabioud B. (1994) Adaptation d'un modèle articulatoire à un autre locuteur, dans le but de contraindre l'inversion articulatori-acoustique. 20e JEP du GFCP de la SFA.
Béland R. (1983) Les modèles du conduit vocal. Travaux de l'Institut de Phonétique de Grenoble, Grenoble. 97 p.
Benguerel A.P. & Cowan H.A. (1974) Coarticulation of upper lip protrusion in French. Phonetica, 30, 41-55.
Benoît C. (1992) The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking faces. J. on Communication, 43, Scientific Society for Telecommunications, Hungary, 32-40.
Benoît C., Lallouache 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 and T.R. Sawallis ed., Elsevier Science Pub., Amsterdam.
Bjuggren G. & Fant (1964) The Nasal Cavity Structures. STL QPSR 4, 5-7.
Boë L.J. & Abry C. (1986) Nomogrammes et systèmes vocaliques. 15e JEP, GCP, GALF, 303-305.
Boë L.-J., Bognar E., Gay T. & Descout R. (1981) Modélisation des mouvements mandibulaires : une optimisation radiale. 4th FASE Symposium, 2, 177-181
Boë L.-J., Maeda S. & Perrier P. (1991) La modélisation du conduit vocal. Le Courrier du CNRS. Dossiers Scientiques, Signaux et images, 64-66.
Boë L.-J. & Perrier P. (1990) Comments on "Distinctive Regions and Modes : A New Theory of Speech Production" by M. Mrayati, R. Carré & B. Guérin. Speech Communication, 9, 217-230.
Boë L.-J., Perrier & Bailly G. (1992) The Geometrical Vocal Tract Variables Controlled for Vowel Production : Proposals for Constraining Acoustic-to-Articulatory Inversion. J. of Phonetics, 20, 27-38.
Boë L.-J., Perrier P., Guérin B. & Schwartz J.-L. (1990) Espace vocalique maximal. EuroSpeech 89, 2, 88-91.
Bognar E. (1983) Espace et cibles mandibulaires. Étude de l'espace articulatoire et des positions cibles de la mâchoire dans les séquences consonne-voyelle chezdeux locuteurs français. Thèse 3e Cycle Phonétique, Université Stendhal, Grenoble.
Bognar E. (1985) Directions des recherches phonétiques consacrées à la mâchoire. Bulletin de l'Institut de Phonétique de Grenoble, 12, 1-45.
Bothe H.H., Rieger F. & Tackman R. (1993) Visual Coarticulation Effects in Syllable Environment. EuroSpeech '93, 3, 1741-1744.
Bothorel A., Simon P., Wioland F. & Zerling J.-P. (1986) Cinéradiographie des voyelles et des consonnes du français. Travaux de l'Institut de Phonétique, Strasbourg.
Branderud P. (1985) Movetrack, a Movement Tracking System. Symposium Franco-Suédois sur la Parole, 2, 113-122.
Brodda B. (1979) Nagot om de Svenska Ordens Fonotax och Morfotax : Iakttagelser med Utgångspunkt från Experiment med Automatisk Morfologisk Analys. Papers from the Institute of Linguistics, Univ. Stockhom, 38, 63-108.
Brooke & Summerfield (1983) Analysis Synthesis and Perception of Visible Articulatory Movements. J. of Phonetics, 11, 63-76.
Browman C. P. & Goldstein L. (1986) Towards an Articulatory Phonology. Phonology Yearbook, 6, 219-252.
Browman C. P. & Goldstein L. (1989) Articulatory Gestures as Phonological units. Phonology, 6, 201-251.
Castelli E. (1989) Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle. Thèse INP, Systèmes électroniques, Grenoble.
Chiba T. & Kajiyama M. (1941) The Vowel, its Nature and Structure. Tokyo Kaseikan Pub., Tokyo.
Coker C.H. (1967) Synthesis by Rule from Articulatory Parameters. Conf. Speech Comm. Process. AFCRL and IEEE Audio and Electracoustic group, Paper A9, 52-53.
Coker C.H. (1968) Speech Synthesis with a Parametric Articulatory Model. Speech Model. Speech Symposium, Kyoto, Paper A-4.
Coker C.H. (1976) A Model of Articulatory Dynamics and Control. Proc. IEEE 64, 452-460.
Coker C.H. & Fujimura O. (1966) A Model for Specification of Vocal Tract Area Function. J. Acoust. Soc. Am., 40, 1271.
Coker C.H., Umeda N. & Browman C.P. (1973) Automatic Synthesis from Ordinary English Text. In *Speech Synthesis*, J.L. Flanagan & L.R. Rabiner ed., 400-411, Dowden, Hutchinson & Ross, Stroudsbourg.
Dagognet F. (1987) Etienne-Jules Marey. Hazan, Collection 35/37, Paris.
Dart S.N. (1987) A Bibliography of X-Ray Studies of Speech. UCLA Papers in Phonetics, 66, 1-97.
Demolin D. & Segebarth C. (1992) Analyse de la production des voyelles de quelques langues [du] Soudan Central par IRM. 19e JEP, GFCP, SFA, 37-42.
Dudley H. (1955) Fundamental of Speech Synthesis. J. Audio Engr. Soc., 3, 170-185.
Dunn H.K. (1950) The Calculation of Vowel resonances, and an Electrical Vocal Tract. J. Acoust. Soc. Am., 22, 740-753.
Edwards J. & Harris K.S. (1985) A Two-Dimensional Model of Jaw Movement during Speech. J. Acoust. Soc. Am. 77 S1, S100.
Edwards J. & Harris K.S. (1990) Rotation and Translation of the Jaw during Speech. J. Speech Hearing Research, 33, 550-562.
Fant G. (1960) Acoustic Theory of Speech Production. Mouton, The Hague.
Fant G. (1965) Formants and Cavities. 5th Int. Congr. Phonetic Sciences, 120-140.
Fant G. (1975) The Relations between Area Function and the Acoustic Signal. Phonetica, 37, 55-86.

Fant G. (1992) Vocal Tract Area Functions of Swedish Vowels and a New Three-Parameter Model. ICSLP 92.
Feng G. (1986) Modélisation acoustique et traitement du signal de parole. Le cas des voyelles nasales. Thèse INPG, Automatique et Traitement du Signal, Grenoble.
Flanagan J.L., Coker C.H., Rabiner L.R., Schafer R.W. & Umeda N. (1970) Synthetic Voices for Computers. IEEE Spectrum, 7, 22-45.
Flanagan J., Ishizaka K. & Shipley K.L. (1980) Signal Models for Low-Bit Rate Coding of Speech. J. Acoust. Soc. Am., 63, 780-791.
Foldvik A.K., Husby O., Kværness J., Nordli I.C. & Rinck P.A. (1991) MRI (magnetic Resonance Imaging) for Filming Articulatory Movements. XII^{e} Congrès International des Sciences Phonétiques, 2, 34-36.
Foldvik A.K., Kristiansen U. & Kværness J. (1991) A Time-Evolving Three-Dimensional Vocal Tract Model by means of Magnetic Resonance Imagering (MRI). EuroSpeech '93, 1, 557-558.
Fowler C. (1980) Coarticulation and Theories of Extrinsic Timing. J. of Phonetics, 8, 113-133.
Fromkin V. (1964) Lips Position in American English Vowels. Language and Speech, 7, 215-225.
Fujimura O. (1977) Model Studies of Tongue Gestures and the Derivation of Vocal Tract Area Functions. In *Dynamic Aspects of Speech Production*, M. Sawashima & F.S. Cooper ed., 225-232, University of Tokyo Press, Tokyo.
Fujimura O., Ishida H. & Kiritani S. (1968) Computer Controlled Dynamic Cineradiography. Annual Bulletin (Research Institute of Logopedics and Phoniatrics), Univ. of Tokyo, 2, 6-10.
Fujimura O. & Kakita Y. (1978) Remarks on Quantitative Description of the Lingual Articulation. In *Frontiers of Speech Communication Research*, B. Lindblom & S. Öhman ed., 17-24, Academic Press, London.
Fujimura O., Kiritani S. & Ishida H. (1973) Computer Controlled Radiography for Observation of Movements of Articulatory and other Human Organs. Comput. Biol. Med. 3, 371-384, Pergamon Press, Oxford.
Fujisaki H. (1977) Functional Models of Articulatory and Phonatory Dynamics. In *Articulatory Models and Phonetics*. R. Carré, R. Descout & M. Waskop, 127-136. GALF, Groupe de la Communication Parlée.
Gabioud B. (1994) Articulatory Models in Speech Synthesis. In Fundamentals of Speech Synthesis and Speech Recognition. E. Keller ed., J. Wiley, Londres.
Galindo B. & Teston B. (1989-1990) PHYSIOLOGIA : un logiciel d'analyse des paramètres physiologiques de la parole. Travaux de l'Institut de Phonétique d'Aix, 197-217.
Gay T., Lindblom B.E.F. & Lubker J. (1981) Production of Bite-Block Vowels : Acoustic Equivalence by Selective Compensation. J. Acoust. Soc. Am, 69, 802-810.
Gentil M. (1980) Labialité en français. Etude phonétique et aspects physiologiques. Des lèvres : géométrie, cinématique, électromyographie et idiolectologie. Thèse de 3e cycle, Phonétique. Université Stendhal, Grenoble.
Gibbs C.H. & Messerman T. (1972) Jaw Motion during Speech. Am. Speech and Hearing Association, Rep. 7, 104-112.
Goldstein L. (1983) Vowel Shifts and Articulatory-Acoustic Relations. 10th Int. Congr. Phonetic Sciences IIa, 267-273.
Gourret J.P., Paillé J., Boë L.-J. & Descout R. (1985) Système de détection automatique de labiofilms. 14^{e} JEP, GCP, GALF, 61-64.
Guiard-Marigny T. (1992) Animation en temps réel d'un modèle paramétrisé de lèvres. DEA Signal Image Parole, INP, Grenoble.
Hafer E. (1974) Speech Analysis by Articulatory Synthesis. M.S. Thesis Dept. Elec. Eng., Northwestern Univ.
Hardcastle W.J. & Marchal A. (ed.) (1990) Speech Production and Modelling. NATO ASI series, Kluwer Academic Publishers, Dordrecht.
Harshman R., Ladefoged P. & Goldstein L. (1977) Factor Analysis of Tongue Shapes. J. Acoust. Soc. Am. 62, 693-707.
Hashimoto K. & Sasaki K. (1982a) On the Relationship between the Shape and Position of Tongue for Vowels. J. of Phonetics, 10, 291-299.
Hashimoto K. & Sasaki K. (1982b) Performance of Articulatory Parameters to Express Formant Frequencies. J. Acoust. Soc. Japan (E), 3, 141-149.
Hashimoto K. & Tanimoto M. (1978) Approximation of the Tongue Profile by a Quadratic Curve. J. Acoust. Soc. Japan, 34, 140-148.
Hashimoto K. & Suga S. (1986) Estimation of the Muscular Tensions of the Human Tongue by using a Three-Dimensional Model of the Tongue. J. Acoust. Soc. Japan, 7, 39-46. [en anglais].
Hecker M.H.L. (1962) Studies of Nasal Consonants with an Articulatory Speech Synthesizer. J. Acoust. Soc. Am., 34, 179-188.
Heike G. (1977) Ein Comptererzeugtes Artikulationsmodell für das Artikulationstraining. IPKöln – Berichte 8, 9-45.
Heike G. (1979) Articulatory Measurement and Synthesis. Method and Preliminary Results. Phonetica, 36, 294-301.
Heike G. (1980) Artikulatorische Synthese : Beschreibung des Programmpakets ARSDEM. IPKöln – Berichte 8, 9-45.
Hellwag C.F. (1781) De formatione loquelae. Dissert. Tubingen.Traduction française M.P. Monin in Bulletin de la Communication Parlée, 1.
Heinz J.M. & Stevens K.N. (1964) On the Derivation of Area Functions and Acoustic Spectra from Cineradiographic Films of Speech. J. Acoust. Soc. Am. 36, 1037-1038.
Henke W.L. (1966) Dynamic Articulatory Model of Speech production using Computer Simulation. MIT Ph. Diss., Cambridge.
Henke W.L. (1967a) Preliminaries to Speech Synthesis based on an Articulatory Model. IEEE Boston Speech Conf., 170-177.
Henke W.L. (1967b) Preliminaries to Speech Synthesis based on an Articulatory Model. IEEE Conf. Speech Communication and Processing, 170-177.
Hiki S. (1971) A Model of Speech Organs for Speech Synthesis with Physiological Parameters. J. Acoust. Soc. Japan, 27, 435-444.
Hiki S. & Itoh H. (1986) Influence of Palate Shape on Lingual Articulation. Speech Comm., 5, 141-158.
Hiki S. & Harshman R. (1969) Speech Synthesis by Rules with Physiological Parameters. J. Acoust. Soc. Am., 46, 111.
Hiki S. & Niyada K. (1973) Articulatory Model for Vowel Production. In *Speech Processing*, Universty Tokyo Press, Tokyo.
Hiki S., Niyada K. & Oizumi J. (1972) Speech Synthesis by Control of Neurophysiological Parameters. Int. Conf. on Speech Comm. & Proc. paper D3.
Hiki S. & Oizumi J. (1974) Speech Synthesis by Rule from Neurophysiological Parameters. Speech Communication, 1-2, 219-225.
Höhne J., Schönle P., Conrad B., Hong G., Sandner N., Veldscholten H., Appel C. & Wenig P. (1987) Direct Registration of Articulatory Movements versus Acoustic Analysis for Speech Production Modelling and the Treatment of Speech Motor Control Disorders. ICASSP, 1454-1456.
Honda M. (1985) Dynamic Ultrasonic Measurement System of Tongue Shape.Trans. Committee on Speech Res., Acad. Soc. Japan, 84-105.
Honda K. (1993) Modeling Vocal Tract Organs based on MRI and EMG Observations and its Implication on Brain Function. Annual Bulletin RILP (Univ. Tokyo), 27.
Hoole P. & Gfoerer S. (1990) Electromagnetic Articulatography as a Tool in the Study of Lingual Coarticulation. J. Acoust. Soc. Am. 87, S1, S123.

House A.S. (1957) Analog Studies of Nasal Consonants. J. of Speech and Hearing Disorders, 190-204.
House A.S. & Stevens K.N. (1956) Analog Study of the Nazalization of Vowels. J. of Speech and Hearing Disorders, 21, 218-232.
Jackson M.T.T. (1988) Analysis of Tongue Position : Language-Specific and Cross-Linguistic Models. J. Acoust. Soc. Am., 84, 124-143.
Jackson M.T.T. (1988) Phonetic Theory and Cross-Linguistic Variation in Vowel Articulation. UCLA Working Papers in Phonetics, 71, 184 p.
Jankelson B., Sawin C.W., Crane P.F. & Radke J.C. (1975) Kinesiometric Instrumentation : A New technology. J. Am. Dental. Association, 90, 834-840.
Johansson C., Sundberg J., Wilbrand H. & Ytterbergh C. (1983) From Sagittal Distance to Area. Annual Bulletin of the Research Institute of Logopedics and Phoniatrics 12, 1-14, Univ. Tokyo.
Jordan M.I. (1990) Motor Learning and the Degrees of Freedom Problem. In *Attention and Performance*, M. Jeannerod ed., Vol. XIII, Lawrence Erlbaum, Hilsdale, N.J.
Kakita Y. & Fujimura O. (1977) Remarks on Quantitative Description of the Lingual Articulation. In *Frontiers of Speech Communication*, B. Lindblom & S. Öhman eds., 17-24, Academic Press, London.
Kakita Y. & Fujimura O. (1977) Computational Model of the Tongue : A Revised Version. J. Acoust. Soc. Am., 62 S1, S15 (A).
Kakita Y. & Fujimura O. (1984) Mapping from Muscular Contraction Patterns to Formant Patterns in Vowel Space. Trans. of the Comittee on Speech Research, The Acoust. Soc. Japan, S83-100 [In Japanese].
Kakita Y. & Fujimura O. & Honda K. (1985) Computation of Mapping from Muscular Contraction Patterns to Formants Patterns in Vowel Space. In *Phonetic Linguistics. Essays in Honor of Peter Ladefoged*, V. Fromkin, ed., 133-144, Academic Press, London.
Kawato M., Maeda Y., Uno Y. & Suzuki R. (1990) Trajectory Formation of Arm Movement by Cascade Neural Network Model Based on Minimum Change Criterion. Biological Cybernetics, 62, 275-288.
Keller E. & Ostry D.J. (1983) Computerized Measurement of Tongue Dorsum Movement with Pulsed-Echo Ultra-Sound. J. Acoust. Soc. Am., 73, 1309-1315.
Kelso J.A.S., Saltzman E. & Tuller B. (1986) The Dynamical Perspective of Speech Production. J. of Phonetics, 14, 29-59.
Kier W.M. & Smith K.K. (1985) Tongues, Tentacles, and Trunks : The Biomechanics of Movement in Muscular-Hydrostats. Zool. J. Linn. Soc., 83, 307-324.
Kiritani S. (1986) X-Ray Micro-Beam Method for the Measurement of Articulatory Dynamics -Techniques and Results. Speech Communication, 5, 119-140.
Kiritani S., Fujimura O. & Ishida H. (1971) Computer Controlled Radiographic for Observation of Articulatory Movement. 7th Int. Congr. Acoust, 21-C-13.
Kiritani S., Itoh K. & Fujimura O. (1975) Tongue-Pellet Tracking by a Computer-Controlled X-Ray Microbeam System. J. Acoust. Soc. Am., 57, 1516-1520.
Kiritani S., Miyawaki K., Fujimura O. & Miller J.E. (1976) A Computational Model of the Tongue. Ann. Bull. RILP (Tokyo), 10, 243-251.
Kiritani S. & Sekimoto S. (1977) Parameter Description of
Kozhevnikov V.A. & al. (1964) Systems of Transducers and Electronic Devices for Analog Recording and Data Processing in the Study of Speech Articulation. Med. Electron. Biol. Eng. 2, 135-143.
Kröger B. (1993a) A Gestual Production Model and its Application to Reduction in German. Phonetica, 50, 213-233.
Kröger B. (1993b) A Gestual Approach for Controlling an Articulatory Speech Synthesizer. EuroSpeech '93, 3, 1903-1906.
Kuehn D.P. (1973) A Cinefluoraphic Investigation of Articulatory Velocities. PH. D. , Univ. of Iowa.
Laboissière R. (1992) Préliminaire pour une Robotique de la Communication Parlée : inversion et contrôle d'un modèle articulatoire du conduit vocal. Thèse Signal Image Parole, INP, Grenoble.
Laboissière R., Schwartz J.-L. & Bailly G. (1990) Motor Control for Speech Skills : a Connectionist Approach. Connexionist Models Summer School, SanDiego.
Laboissière R., Schwartz J.-L. & Bailly G. (1992) Modélisation du contrôle moteur en production de la parole : vers un « robot parlant ». 5e Colloque de l'ARC.
Ladefoged P., Anthony J.F.K. & Riley C. (1971) Direct Measurements of the Vocal Tract. UCLA Working Papers in Phonetics 19, 4-13.
Ladefoged P. & Bladon A. (1982) Attemps by Human Speakers to Reproduce Fant's Nomograms. Speech Communication 1, 185-198.
Ladefoged P. & Harshman R. (1979) Formant Frequencies and Movements of the Tongue. In *Frontiers of Speech Communication*, B. Lindblom & S. Öhman eds., 25-34, Academic Press, London.
Lallouache M.T. (1990) Un poste "visage-parole". Acquisition et traitement de contours labiaux. 18^{e} JEP du GCP du GALF, 282-286.
Lallouache M.T. (1991) Un poste "Visage-Parole" couleur: Acquisition et traitement automatique des contours des lèvres. Thèse de doctorat INP Grenoble.
Lallouache M.T. & Abry C. (1992) Une autre vision des modèles d'anticipation. 19^{e} JEP du GFCP de la SFA, 47-52.
Lallouache M.T. & Teston B. (1991) Les signaux de la parole ne sont pas que du son. Le Courrier du CNRS. Dossiers Scientiques, Signaux et images, 15-16.
Lallouache M.T. & Worley C. (1988) Saisie, édition, et traitement d'images et de signaux articulatoires lèvres et mâchoire. Journal d'Acoustique 1, 215-220.
Larar J.N., Schroeter J. & Sondhi N.M. (1988) Vector Quantization of the Articulatory Space. IEEE Trans. on Acoustics, Speech , and Signal Processing 36, 12, 1812-1818.
Le Goff B. (1993) Commande paramétrique d'un modèle de visage 3D pour animation en temps réel. DEA Signal Image Parole, INP, Grenoble.
Leipp E. (1967) Structure physique et contenu sémantique de la parole. Revue d'Acoustique, 3-4, 259-266.
Liljencrants J. (1971) A Fourier Series Description of the Tongue Profile. STL-QPSR, 4, 9-18.
Lin Q. (1990) Speech Production Theory and Articulatory Speech Synthesis. Thesis, RIT, Stockholm.
Lindblom B.E.F. (1972) Numerical Models in the Study of Speech Production and Speech Perception : some Phonological Implications. 7th Int. Congr. Phonetic Sciences, 63-93.
Lindblom B.E.F. (1977) Some uses of Articulatory Models in Phonetic Theory. In *Articulatory Models and Phonetics*. R. Carré, R. Descout & M. Wajskop, 99-104. GALF, Groupe de la Communication Parlée.
Lindblom B.E.F. (1983) Economy in Speech Gesture. In The Production of Speech, P.F. Mac Neilage ed., 217-245, Springer-Verlag, New-York.
Lindblom B.E.F. (1990) In *Speech Production and Speech Modelling*, W.J. Hardcastle & A. Marchal ed., 131-150, Kluwer, Dordrecht.
Lindblom B.E.F. & Sundberg J.E.F. (1971) Acoustical Consequences of Lip, Tongue and Jaw Movements. J. Acoust. Soc. Am., 50, 4, 1166-1179.
Lindblom B.E.F. & Sundberg J. (1974) Modeling Coarticulation in Apical Stops. Speech Communication Seminar, 1-2, 87-94.
Lindqvist J. & Sundberg J. (1971) Pharyngeal Constrictions. STL QPSR 4, 26-31.

Linker W. (1982) Articulatory and Acoustic Correlates of Labial Activity in Vowels : A Cross-Linguistic Study.UCLA Working Papers in Phonetics, 56, 134 p.
Lœvenbruck H. & Perrier P. (1993) Vocalic Reduction : Prediction of Acoustic and Articulatory Variabilities with Invariant Motor Commands. EuroSpeech '93, 1, 85-88.
Lonchamp F. (1988) Études sur la production de la parole. Les indices acoustiques de la nasalité vocalique. La modification du timbre par la fréquence fondamentale. Doctorat d'Etat en Lettres et Sciences Humaines. Nancy.
Louvion J.R. (1980) Détection automatique du contour des lèvres et extraction des paramètres constitutifs. Thèse 3e Cycle, Informatique. Univ. Rennes.
McCutcheon M.J., Fletcher S.G. & Hasegawa A. (1977) Video-Scanning System for Measurement of Lip and Jaw Motion. J. Acoust. Soc. Am., 53-1070-1082.
McGowan R.S. (1994) Recovering Articulatory Movement from Formant Frequency Trajectories using Task Dynamics and a Genetic Algoritm. Speech Communication,
Mac Neilage P., Ladefoged P & Sawashima M. (1979) Report : Speech Production. 9th Int. Congr. Phonetic Sciences, 3, 29-40.
Maeda S. (1972) On the Conversion of Vocal Tract X-Ray Data into Formant Frequencies. Bell Laboratories, Murray Hill.
Maeda S. (1978) Une analyse statistique sur les positions de la langue : étude préliminaire sur les voyelles françaises. 9e JEP, GCP, GALF, 191-199.
Maeda S. (1979a) An Articulatory Model of the Tongue based on a Statitical Analysis. J. Acoust. Soc. Am. 65, S22.
Maeda S. (1979b) Un modèle articulatoire de la langue avec des composantes linéaires. 10e JEP, GCP, GALF, 152-162.
Maeda S. (1982a) The Role of the Sinus Cavities in the Production of Nasal Vowels. ICASSP IEEE, 911-914.
Maeda S. (1989) Articulation compensatoire des voyelles : Analyse de données cinéradiographiques avec un modèle linéaire. Mélanges Péla Simon, 2, 545-562.
Maeda S. (1990) Compensatory Articulation during Speech : Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In *Speech Production and Speech Modelling*, W.J. Hardcastle & A. Marchal ed., 131-150, Kluwer, Dordrecht.
Maeda S. (1991) On Articulatory and Acoustic Variability. J. of Phonetics, 19, 321-331.
Maeda S., Bothorel A., Simon P., Wioland F. & Zerling J.P. (1993) Software for Digitazation of Labial and X-Ray Film Data. Sound-to-Gesture Inversion in Speech. An ESPRIT Basic Research Project n° 6975.Year 1 report, III, WP2.1, SW1.
Maeda S. & Honda K. (1994) From EMG to Formant Patterns of Vowels : the Implication of Vowel Systems Spaces. Phonetica, à paraître.
Majid R. (1986) Modélisation articulatoire du conduit vocal. Exploration et exploitation. Fonctions de macro-sensibilité paramétriques et voyelles du français. Thèse Docteur Ingénieur, INP Grenoble.
Majid R., Abry C., Boë L.-J. & Perrier P. (1987) Contribution à la classification articulatori-acoustique des voyelles : étude des « macrosensibilités » à l'aide d'un modèle articulatoire. 11th Int. Congr. Phonetic Sciences, 2, 348-351.
Marchal A. (1985) L'électropalatographie : contribution à l'étude de la coarticulation dans les groupes d'occlusives. Thèse d'Etat, Phonétique, Nancy II, 1070p.
Marchal A. (1988) La palatographie. Editions du CNRS, Paris.
Marey E.-J. (1878) La méthode graphique dans les sciences expérimentales et particulièrement en physiologie et en médecine. Paris.
Mermelstein P. (1973) Articulatory Model for the Study of Speech Production. J. Acoust. Soc. Am., 53, 1070-1082.
Minifie F.D., Hixon T.J., Kelsey C.A. & Woodhouse R. (1970) Lateral Pharyngeal Wall Movement during Speech Production. J. Speech Hearing Research 13, 584.
Mohamadi T. (1993) Synthèse à partir du texte de visages parlants : réalisation d'un prototype et mesures d'intelligibilité bimodale. Doctorat INP, Signal Image Parole, Grenoble.
Mokbel C. (1989) Étude et modélisations articulatoires. Liens entre fonction d'aire et modèle à huit régions. DEA Systèmes Électroniques, INP, Grenoble.
Moll K.L. (1960) Cinefluorographic Techniques in Speech Research. J. Speech and Hearing Research 3, 277.
Moll K.L. & Daniloff R.G. (1970) Investigation of the Timing of Velars Movements during Speech. J. Acoust. Soc. Am. 50, 678-684.
Mrayati M. (1974) Contribution aux études sur la production de la parole. Modèles électriques du conduit vocal avec pertes, du conduit nasal et de la source vocale. Étude de leurs interactions. Relations entre disposition articulatoire et caractéristiques acoustiques. Doctorat d'État, INP Grenoble.
Mrayati M., Carré R. & Guérin B. (1988) Distinctive Regions and modes : A New Theory of Speech Production. Speech Communication 7, 257-286.
Muller E.M. (1975) Biomechanics of Lip Movement for Speech Production. Unpublished manuscript. Univ. Washington.
Muller E.M. & Abbs J.H. (1979) Strain Gauge Transduction of Lip and Jaw Motion in the Midsagittal plane Refinement of a Prototype System. J. Acoust. Soc. Am., 65, 481-486.
Nakajima T. (1977) Identification of a Dynamic Articulatory Model by Acoustic Analysis. In In *Dynamic Aspects of Speech Production*, M. Sawashima & F.S. Cooper ed., 251-275, University of Tokyo Press, Tokyo.
Nakajima T., Ohmura N. & Ishizaki S. (1975) Linear Articulatory Model and its Applications. Progress Report on Speech Research, PIPS-R, 11, Electrotechnical Lab. Tokyo.
Nakajima T., Omura N., Tanaka K. & Ishizaki S. (1974) Estimation of Vocal Tract Area Functions by Adaptative Inverse Filtering Methods and Identification of Articulatory Model. Speech Communication Seminar 1, 11-20.
Niimi S. & Shimada J. (1984) Ultrasonic Observation of Tongue Dynamics. Ann. Bull. Res. Inst. Log. Phoniat. Univ. of Tokyo, 18, 13-17.
Ohala J., Hiki S., Hubler S. & Harshman R. (1968) Photo-Electric Methods of Transducing Lip and Jaw Movements in Speech. UCLA Working Papers 10, 135-144.
Öhman S.E.G (1964) Numerical Model for Coarticulation using a Computer-Simulated Vocal Tract. J. Acoust. Soc. Am., 36, 1038.
Öhman S.E.G (1966) Coarticulation in VCV Utterances Spectrographic Measurements. J. Acoust. Soc. Am., 40, 979-988.
Öhman S.E.G (1967) Numerical Model of Coarticulation. J. Acoust. Soc. Am., 41, 310-320.
Ostry D., Keller E. & Parush A. (1983) Similarities in the Control of Speech Articulators and the Limbs : Kinematics of Tongue Dorsum Movement in Speech. J. Exp. Psychology : Human Perception and Performance, 9, 622-636.
Ostry D.J. & Munhall K.G. (à paraître) Control of Jaw Orientation and Position in Mastication and Speech. Submitted to J. of Neurophysiology.
Palacios E., Fine M. & Haughton V.M. (1980) Multiplanar Anatomy of the Head and the Neck. A Wiley Medical Pub., John Wiley & Sons, New York.
Panagos G. & Strube H.W. (1987) Two Magnetic Methods for Articulatory Measurements. Phonetica, 44, 197-209.

Perkell J. S. (1969) Physiology of Speech Production. Results and Implications of a Quantitive Cineradiographic Study. MIT Press, Cambrige Mass.
Perkel J. S. (1974) A Physiologically-Oriented Model of Tongue Activity during Speech Production. Ph. Diss., MIT, Cambridge.
Perkell J.S., Cohen M., Svirksy M., Matthies M., Garabieta I. & Jackson M. (1992) Electromagnetic Midsagittal Articulometer (EMMA) Systems for Transducing Speech Articulatory Movements. J. Acoust. Soc. Am., 92, 3078-3096.
Perkell J.S. & Matthies M.L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/ : within- and cross-subject variability. J. Acoust. Soc. Am., 91, 2911-2925.
Perrier P. (1991) De l'usage de modèles pour l'étude de la production de la parole. Habilitation à Diriger des Recherches. INP, Grenoble.
Perrier P., Abry C. & Keller E. (1988) Vers une modélisation des mouvements du dos de la langue. J. d'Acoustique 2, 69-77.
Perrier P., Badin P. & Boë L.-J. (1987) Nomogrammes du conduit vocal par modélisation articulatoire. 16^{e} JEP, GCP, GALF, 124-127.
Perrier P. & Boë L.J. (1989) Passage de la coupe sagittale à la fonction d'aire. J. Acoustique, 2, 59-67.
Perrier P., Boë L.-J. & Sock R. (1992) Vocal Tract Area Function Estimation from Midsagittal Dimensions with CT Scans and a Vocal Tract : Modeling the Transition with two Sets of Coefficients. J. of Speech and Hearing Research, 35, 53-67.
Purves B., Blackett K. & Strong W. (1970) Speech Synthesis with a Vocal tract Synthesizer. J. Acoust. Soc. Am., 47, 93 (A).
Robert-Ribes J., Lallouache T., Escudier P. & Schwartz J.-L. (1993). Integrating Auditory and Visual Representations for Audiovisual Vowel Recognition. EuroSpeech '93, 3, 1753-1756.
Rousselot P. L'abbé (1897-1908) Principes de phonétique expérimentale. (2 vol.), Welter, Paris.
Rouy V. (1992) Suivi des mouvements visibles de la parole en vidéo rapide. DEA Signal Image Parole, INP, Grenoble.
Rubin P.E., Baer T. & Mermelstein P. (1981) An Articulatory Synthesizer for Perceptual research. J. Acoust. Soc. Am., 70-321-328.
Saltzman E.L. & Munhal K.G. (1989) A Dynamical Approach to Gestural Patterning in Speech Production. Ecological Psychology, 1, 1615-1623.
Sanchez H. & Boë L.-J. (1984) De la coupe sagittale à la fonction d'aire. Bull. Institut de Phonétique de Grenoble, 13, 1-24.
Shadle C.H., Carter J.N., Monks T.P. & Field J. (1993) Depth Measurement of Face and Palate by Structured Light. EuroSpeech '93, 3, 1745-1748.
Shigenawa M. & Arizumi H. (1977) Articulatory Movements by Rule. In *Articulatory Models and Phonetics*. R. Carré, R. Descout & M. Waskop, 193-202. GALF, Groupe de la Communication Parlée.
Shirai K. & Honda M. (1977) An Articulatory Model and the Estimation of Articulatory Parameters by Nonlinear Regression Model. Trans. IECE Japan, 59-A, 8, 668-674.
Shirai K. & Honda M. (1977) Estimation of Articulatory Motion. In *Dynamic Aspects of Speech Production*, M. Sawashima & F.S. Cooper ed., 279-302, University of Tokyo Press, Tokyo.
Simon P. (1967) Film radiologique des articulations et les aspects génétiques des sons du langage. Orbis, 10, 1.
Sonoda Y. (1974) Observation of Tongue Movements employing Magnetometer Sensor. IEEE Trans. Magn., MAG-10, 954-957.
Sonoda Y. (1983) Use of an Optical Position-Sensitive Device for Recording Lip and Jaw Movements in Speech. 10th Int. Congr. Phonetic Sciences, IIb, 378-382.
Sonoda Y. & Wanishi S. (1982) New Optical Method for Recording Lip and Jaw Movements. J. Acoust. Soc. Am., 72, 700-704.
Sonies B.C., Shawker T.H., Hall T.E., Gerber L.H. & Leighton S.B. (1981) Ultrasonic Vizualization of Tongue Motion during Speech. J. Acoust. Soc. Am., 70, 683-686.
Sundberg (1969) On the Problem of Obtaining Area Function from Lateral X-Ray Pictures of the Vocal-Tract. STL QPSR 1, 43-45.
Sussman H.M. & Smith K.U. (1970) Transducer for Measuring Mandibular Movements during Speech. J. Acoust. Soc. Am., 48, 857-858.
Stetson R.H. (1928) Motor Phonetics. A Study of Speech Movements in Action. Archives Néerlandaises de Phonétique Expérimentale 3, 1-216. 2nd ed. 1951, North Holland Publishing Co., Amsterdam. A Retropective Edition, 1987, J.A.S. Kelso & K.G. Munhall ed., A College-Hill Pub., Boston.
Stevens K.N. (1989) On the Quantal nature of Speech. J. Phonetics, 17, 3-45.
Stevens K.N. (1972) The Quantal Nature of Speech : Evidence from Articulatory-Acoustic Data. *In Human Communication : A Unified View*. E.E. David Jr. & P.B. Denes ed., 51-66. Mc Graw Hill, New-York.
Stevens K.N. & House A.S. (1955) Development of a Quantitative Description of a Vowel Articulation. J. Acoust. Soc. Am., 27-484-493.
Takeuchi S., Kasuya H. & Kido K. (1977) A Study of the Effects of Nasal and Paranasal Cavities on the Spectra of Speech Sounds. J. Acoust. Soc. Japan [en Japonais] 33, 163-172.
Vallée N. & Boë L.J. (1992) Vers des prototypes acoustiques et articulatoires des 37 phonèmes vocaliques d'UPSID. 19^{e} JEP, GCP, GALF, 53-58.
Wilhems-Tricario R. (à paraître) Methods for Modelling the Soft-Tissue Speech Articulators. ATR Technical Report.
Wioland F. (1971a) Les mouvements du maxillaire dans la chaîne parlée. Trav. de l'Institut de Phonétique de Strasbourg, 3, 57-119.
Wioland F. (1971b) Représentation continue sous forme de courbes des mouvements du maxillaire. Thèse de 3^{e} Cycle, Phonétique, Strasbourg.
Wood S. (1982) X-Ray and Model Studies of Vowel Articulation. Lund Univ. Ph. D. Working Papers, Lund Univ. Dept. Linguistics and Phonetics 23.
Zemlin W.R. (1968) Speech and Hearing Science. Anatomy and Physiology. Prentice Hall, Englewood Cliffs, New Jersey.
Zerling J.-P. (1979) Articulation et coarticulation dans les groupes occlusive-voyelle en français. Thèse de 3^{e} Cycle, Phonétique, Nancy.
Zerling J.-P. (1990) Aspects articulatoires à partir de labiophotographies, labiofilms et films radiologiques. Etude statique, dynamique et constrastive. Thèse d'Etat, Phonétique, Strasbourg.
Zohair L. (1993) Constitution d'une banque de données pour la synthèse articulatori-acoustique. Thèse Sciences du Langage, Université Stendhal.

Remerciements

Ils s'adressent à Christian Abry, Ali Adjoudani, Gérard Bailly, Rafael Laboissière, Thierry Guiard-Marigny, pour leurs suggestions et documents.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O2

PROSODIE

PATRONS PROSODIQUES ET INTENTIONS DES LOCUTEURS: PRODUCTION ET PERCEPTION DE FORMES EXPRESSIVES CHEZ L'ADULTE ET L'ENFANT.

CLAIRE GERARD ET JULIETTE CLEMENT

LABORATOIRE DE PSYCHOLOGIE EXPÉRIMENTALE
UNIVERSITÉ RENÉ DESCARTES, URA CNRS 316

Résumé

Divers déterminants pragmatiques influent sur les mécanismes de production, perception et compréhension du langage et le ruban sonore est susceptible de les refléter. Comment s'élabore avec l'âge la prise en compte de ces déterminants ? Toute communication suppose par exemple une intention particulière accompagnant le contenu du message: simple information, question, mise en doute, demande de confirmation, ou encore approbation, regret, admiration...La première partie de notre étude porte sur la production des intentions: elle cherche à mieux cerner les régulations temporelles et intonatives du discours lors de l'expression d'un même énoncé par les mêmes locuteurs dans des intentions différentes. Les expressions affirmative et interrogative de ces énoncés réalisés par 20 locuteurs (12 adultes des deux sexes et 8 enfants de 7 ans) sont comparées aux expressions des mêmes phrases mais manifestant la joie ou le regret. La seconde partie de l'étude porte sur la perception des intentions. A cette fin, 90 auditeurs (72 sujets de 6, 8, et 10 ans et un groupe de 18 adultes) doivent identifier les intentions exprimées par 3 types de locuteurs différents (femme, homme et enfant) lors de l'énonciation de trente six phrases extraites du corpus analysé. Lors de la phase de production, les valeurs moyennes de Fo, les contours mélodiques et les durées d'énonciation ont été mesurés au niveau global de la phrase entière et au niveau plus local, mot par mot, grâce à un éditeur du signal sonore. Ces mesures permettent de constater qu'adultes et enfants de 7 ans organisent leurs énoncés de façon similaire et stable quand à la différenciation des intentions. Mais la détection des intentions lors de la deuxième phase de l'étude est plus difficile chez les enfants même jusqu'à 10 ans (par rapport aux scores des adultes) ce qui confirmerait l'existence d'un décalage entre production et perception de la prosodie dans le développement des mécanismes de compréhension de la parole

Mots-clés: *prosodie, intentions expressives, acquisition de la parole.*

INTRODUCTION

La production, la perception et la compréhension du langage sont envisagées dans le cadre de la psychologie cognitive comme des opérations portant sur des représentations mentales. Les formes prosodiques témoignant des intentions des locuteurs pourraient faire partie de ces représentations mentales et servir de formes de référence auxquelles sont comparées les réalisations sonores actuelles d'un énoncé donné (Sorin, 1989). Si ces représentations mentales gouvernent réellement perception et production du discours, des invariants structuraux doivent pouvoir être mis en évidence malgré la variabilité intra- et inter-locuteurs. Notre étude cherche donc à cerner la nature de ces invariants, mais aussi à préciser, dans une perspective développementale, d'une part l'émergence chez l'enfant de configurations prosodiques obéissant au modèle adulte lors de la production, et d'autre part le développement du "décodage" des intentions lors de la perception de la parole. Sur la base des études de Cutler & Swinney (1987) nous avons fait l'hypothèse d'un décalage entre maîtrise de la production et maîtrise de la perception de la prosodie et cherchons à en préciser la nature en examinant une tranche d'âge considérée comme critique de ce point de vue par divers auteurs (Moore, Harris & Patriquin (1993), Winner & Leekam (1991)). Les recherches sur l'acquisition des formes prosodiques montrent que celle-ci s'effectue en plusieurs étapes (depuis l'imitation des contours mélodiques de la voix de la mère lors des premières vocalises (Crystal, 1986) jusqu'à l'identification précise des intentions et expressions des locuteurs), et surtout suggèrent l'existence du décalage suivant entre production et perception: alors que les productions sonores des enfants manifestent précocément les mêmes configurations prosodiques que la parole des adultes, la compréhension d'intentions différentes marquées par des formes prosodiques particulières semble tardive. Notre propos n'est pas de chercher à mieux connaître les étapes de l'acquisition de la parole, mais d'utiliser ce décalage (s'il se vérifie) pour mieux cerner la nature des formes prosodiques fondamentales (précoces, gouvernant à la fois production et perception) et la

nature des formes prosodiques moins stables, ou plus subtiles (tardives, difficiles à identifier). Ceci permettrait peut-être d'éclairer à la fois la nature et la fonction de ces représentations mentales hypothétiques censées gouverner le codage et le décodage de la parole. Les paramètres prosodiques (intensité, durée, fréquence fondamentale) s'associent généralement en configurations prosodiques globales. Dans une étude antérieure portant sur des adultes (Gérard & Rigaut, 1994) nous avions montré que les contours mélodiques classiques des affirmations et interrogations (respectivement descendant et montant) étaient systématiquement associés à des rythmes différents (rapide pour l'interrogation, plus lent pour l'affirmation). Chez les enfants, Crystal (1986) a montré que les contours typiques de ces deux types d'énoncés apparaîssent précocément, mais qu'en est-il des durées d'énonciation? De même notre étude de l'expression de la joie et du regret montrait chez les adultes une configuration particulière des durées et des fréquences fondamentales. Nous avons donc désiré rechercher ces configurations globales chez l'enfant en reprenant la même procédure que celle utilisée auparavant: l'expression répétitive d'un même énoncé (une phrase) selon des intentions différentes. Cette procédure permet 1. la vérification de la stabilité des formes prosodiques au travers des répétitions pour un même locuteur; 2. la comparaison d'intentions différentes pour un même contenu articulatoire, lexical, sémantique. La comparaison adulte/enfant devrait permettre d'affiner le raisonnement sur le décalage production/perception chez l'enfant, en prenant en considération des indices plus fins concernant la production d'une part, et en testant les mêmes énoncés lors d'une phase perceptive ensuite.

METHODE

Lors de la tâche de **production** pour les adultes, trois phrases de huit syllabes, toutes composées d'un sujet, d'un verbe et d'un complément d'objet ou d'un complément circonstanciel ont été choisies : (P1) "Ils ont découvert le secret"; (P2) "Ils repartent à Londres mardi"; (P3) "On emmène Michel en vacances". Les phrases sont précédées de "mots introducteurs" tirés du langage courant et facilitant l'expression de cinq intentions : *affirmation* introduite par "Tu sais"; *interrogation* introduite par "Quoi ?"; *joie* introduite par "Chouette"; *regret* introduit par "Zut"; *admiration* introduite par "Oh". Les 12 locuteurs adultes (6 hommes et 6 femmes) sont des étudiants en psychologie ne présentant pas de trouble auditif ni articulatoire. L'enregistrement s'est fait en chambre sourde. Les locuteurs prenaient préalablement connaissance des mots introducteurs et des phrases, inscrits sur des feuilles différentes, pendant quelques minutes. Pendant l'enregistrement, l'expérimentateur présentait chacune des phrases l'une après l'autre et indiquait oralement l'intention dans laquelle la phrase (que le locuteur gardait sous ses yeux) devait être dite, en même temps qu'il présentait au-dessus de celle-ci le mot introducteur de cette intention. Les locuteurs ont énoncé ainsi chaque phrase dans les cinq intentions avant de passer à la phrase suivante. Après une courte pause, les locuteurs ont effectué exactement la même tâche une seconde, puis une troisième fois. L'ordre de présentation des 3 phrases P1 à P3 et l'ordre de présentation des intentions pour chaque phrase étaient contrebalancés sur l'ensemble des 12 sujets. Soient S les locuteurs, X leur sexe, P les 3 phrases, I les 5 intentions, R les trois répétitions, le plan de l'expérience s'écrit: S6<X2>*P3*I5*R3. Les 8 enfants de 7 ans étudiés ensuite ont effectué une tâche simplifiée (le plan d'analyse devient S8*I4*R3): une seule phrase (P3) devait être prononcée selon 4 intentions (la forme admirative a été supprimée) après familiarisation à la tâche grâce à une phrase d'entraînement. Cette épreuve simplifiée est répétée trois fois également. Après analyse grâce à un éditeur du signal sonore (Unice, Vecsys) des 636 énoncés ainsi recueillis, les indices suivants ont fait l'objet d'analyses de la variance : 1 - indices "globaux" calculés sur l'ensemble de la phrase = durée totale, Fo moyenne, contour mélodique (indice de Cooper & Eady), valeur du pic de Fo, gamme de variations de Fo. 2 - indices "locaux" = durée et Fo moyenne de chaque mot de chaque phrase.

Pour la tâche de **perception**, des énoncés correspondant aux quatre intentions de base (affirmation, interrogation, joie, regret) ont été sélectionnés dans le corpus analysé précédemment, en réintégrant les 3 phrases différentes P1, P2 et P3. Après ablation du mot introducteur pour supprimer tout indice sémantique caractérisant les intentions, ces phrases sont présentées aux auditeurs qui doivent les affecter, par choix forcé, à l'une des quatre intentions. Les énoncés produits par trois types de locuteurs (enfant, femme et homme, sélectionnés par les expérimentateurs en raison de leurs qualités articulatoires aussi bien qu'expressives au sein des populations précédemment étudiées) sont présentés selon des ordres contrebalancés quant au contenu sémantique des 3 phrases et aux intentions, mais regroupés par locuteur pour favoriser une habituation à un type de voix. Les 24 auditeurs pour chaque groupe d'âge (6 ans, 8 ans et 10 ans) tracent une croix dans l'une des 4 cases d'un tableau où chacune des cases est surmontée d'un dessin représentatif des intentions (visages "simplifiés" neutre, gai, triste, et point d'interrogation). Les auditeurs adultes, au nombre de 18, font exactement la même chose et seront considérés comme un groupe de référence auquel comparer les résultats des enfants. Le plan de l'expérience est alors

le suivant: S<A4>*P3*I4*L3 où S désigne les sujets, A les âges des auditeurs (6, 8 et 10 ans et les adultes), P les trois types de phrases, L les trois types de locuteurs et I les quatre intentions sélectionnées. Au total, chacun des 90 auditeurs doit donc identifier l'intention exprimée lors de l'écoute de 36 phrases.

RESULTATS

Production des intentions: Pour les locuteurs adultes, quel que soit l'indice mesuré, global ou local, la répétition de la tâche n'entraîne jamais de différence significative, et ne s'exprime jamais non plus sous forme d'interaction significative avec aucun des autres facteurs (groupe de sujets, type de phrase, type d'intention exprimée). Donc les indices globaux et locaux mesurés "résistent" à la variabilité intra-locuteurs lors de la répétition de l'énoncé à un moment ultérieur. Le deuxième résultat essentiel est que, toujours par rapport aux indices que nous mesurons, la variabilité inter-locuteurs est également faible. Ainsi par exemple, le facteur "groupe de sujets" (qui renvoie au sexe des sujets) se marque bien sûr par des différences des valeurs de Fo brutes (moyennes, ou pics) mais ne joue ni sur les contours, ni sur les durées. Aucune interaction avec les autres facteurs (phrases, intentions, répétitions) n'étant en outre observée, on peut conclure que cette variabilité inter-locuteurs s'exerce sans doute plus au niveau segmental que suprasegmental. Le seul et unique facteur significatif (qui donc l'emporte sur les différences inter-individuelles) est de façon systématique le type d'intention exprimée, et ce facteur conduit à des différences significatives répétées et constantes quelque soit le contenu lexical et sémantique de la phrase. Il en est de même chez les enfants. Des analyses "locales" ont été effectuées ensuite qui avaient pour but de "cibler" les marques de l'intentionnalité, en recherchant si l'un ou l'autre des mots pleins de la phrase (noms, verbes) pouvait à lui seul être porteur des modifications de durée ou de fréquence enregistrées. En ce qui concerne les adultes, les résultats de ces analyses ont été exposés ailleurs (Gérard & Rigaut, 1994). Ils nous ont permis de développer des hypothèses spécifiques sur l'accumulation progressive des marques prosodiques au cours du temps. En ce qui concerne les enfants, ces analyses locales confirment les conclusions tirées à propos des adultes, et soulignent que c'est surtout en fin de phrase que les différences dues aux intentions sont les plus marquées. Dans le but de nous focaliser sur les comparaisons adultes/enfants qui forment le coeur de la nouvelle série d'expériences présentée ici, seuls les indices globaux principaux issus d'une partie des données vont être pris en considération.

Le tableau 1 présente les valeurs moyennes de durée et fréquence fondamentale obtenues pour la phrase 3 (seule phrase commune à la population des adultes et à celle des enfants) lors de l'énonciation des quatre intentions par les trois groupes de locuteurs hommes, femmes et enfants. Les analyses de variance effectuées sur les données des adultes montraient que ce qui est observé pour la phrase donnée en exemple dans le tableau ci-dessous s'observe de façon strictement identique pour les autres phrases. Les analyses de variance effectuées sur les données des enfants montrent des différences significatives, analogues à celles des adultes, quand l'intention change, aussi bien pour les valeurs moyennes de F0 que pour les durées.

Tableau 1: *Durées en msec (D) et Fo en Hz moyennes pour chacun des groupes (G1: enfants, G2: femmes, G3: hommes) lors de l'énonciation de la phrase " On emmène Michel en vacances" selon quatre intentions (I1: affirmation, I2: interrogation, I3: joie, I4: regret).*

		Intentions			
		I1	**I2**	**I3**	**I4**
G1	D	1716	1585	1648	1772
	F0	270	294	279	247
G2	D	1341	1290	1320	1391
	F0	225	252	232	210
G3	D	1271	1206	1271	1323
	F0	126	165	147	114

En ce qui concerne la durée d'énonciation de la phrase, une simple ordination des valeurs observées pour chaque intention montre que les données s'organisent de manière identique pour les trois groupes de locuteurs. Hommes et femmes ne diffèrent pas de façon statistiquement significative quant à leur rythme, ni globalement, ni par intention, alors que les enfants parlent systématiquement plus lentement, mais ces enfants varient de la même façon que les adultes leur rythme d'énonciation lorsqu'ils changent d'intention. La hiérarchie des durées est la suivante:

- pour G1: I2<I3<I1<I4
- pour G2: I2<I3<I1<I4
- pour G3: I2<I3=I1<I4

En ce qui concerne la hauteur moyenne de la voix, on ne s'étonnera pas d'observer les voix les plus aigües chez les enfants, puis chez les femmes puis les voix graves des hommes, avec respectivement des différences moyennes de 47 Hz (soit 17.2 % si l'on ramène cette différence à la hauteur moyenne de la voix), 42 Hz (18.3 %) et 51 Hz (37 %) entre l'intention exprimée sur le ton le plus grave et l'intention exprimée sur le ton le plus aigu. La différenciation des intonations semble donc moindre chez les enfants et les femmes. Mais on peut constater une identité parfaite de la relation d'ordre (du plus

grave au plus aigu) établie en fonction des intentions dans les 3 groupes de locuteurs: I4<I1<I3<I2 pour G1, G2 et G3. Les valeurs des contours (indice de Cooper & Eady calculé sur la phrase entière) sont aussi très semblables dans les trois groupes, avec des pentes marquées (et respectivement descendantes et montantes) pour I1 et I2, ambigües pour I3 et I4. Dès l'âge de 7 ans, des configurations prosodiques globales analogues au "modèle" adulte en durée et fréquence sont donc observées lors de la production des 4 intentions expressives sélectionnées.

Perception des intentions: Peut-on maintenant déceler des comportements perceptifs divergeants en fonction de l'âge des auditeurs et du type d'intention exprimé par les locuteurs ? Les résultats essentiels concernent l'évolution des performances avec l'âge et leur différenciation en fonction des intentions. Ces deux facteurs (âge et intentions) entraînent des différences statistiquement significatives. Le tableau 2 présente les scores de "bonne détection" des intentions enregistrés chez les auditeurs.

Tableau 2: *Pourcentages de bonnes réponses dans la détection des quatre intentions (I1: affirmation, I2: interrogation, I3: joie, I4: regret) pour chaque groupe de locuteurs (G1: enfants de 6 ans, G2: enfants de 8 ans , G3: enfants de 10 ans, G4: adultes).*

	Intentions			
	I1	**I2**	**I3**	**I4**
G1	49.1	23.6	43.5	42.1
G2	67.1	59.3	58.3	56.9
G3	81.9	81.5	66.7	60.6
G4	88.3	99.4	66.7	72.8

A 6 ans les bonnes détections sont minoritaires (moyenne générale = 39.6 %) mais supérieures à une répartition des réponses au hasard (qui produirait 25 % de bonnes réponses pour chaque intention). L'amélioration du décodage des intentions est nette à 8 ans (moyenne = 60.4 %, donc amélioration = 20,8 %), se ralentit à 10 ans (amélioration 12,3, moyenne générale =72.7 %). Les adultes détectent en moyenne 81.8 % des intentions exprimées. Mais ces progrès n'ont pas la même ampleur selon le type d'intention. Les deux expressions de référence (affirmation et interrogation) atteignent de forts scores dès 10 ans alors que les intentions I3 et I4 restent mal détectées même chez les adultes. Ces comportements peuvent être éclairés en précisant la nature des confusions perceptives. En effet, l'incertitude face à une intention donnée peut conduire à divers types de réponses: choix au hasard parmi les 4 réponses possibles, choix de l'intention neutre comme "valeur-refuge", réelle confusion avec une autre intention en raison de la similitude des contours ou des durées, etc... En raisonnant à l'inverse, on peut dire que si une certaine représentation mentale de telle ou telle intention guide la détection, alors certains types d'intention pourraient être difficiles à identifier dans le cas où les locuteurs la réalisent de façon ambigüe, mais ne seront jamais choisis par erreur ou ne seront pas des "valeurs-refuge" lors d'une incertitude. Ce raisonnement s'applique en particulier aux représentations "précoces", comme l'interrogation. Le tableau 3 ci-après permet de préciser la répartition des erreurs commises lors des choix erronés, et les regroupements en ligne montrent que les choix erronés ne se distribuent pas de façon égale sur n'importe quel type d'intention, et que cette différenciation se manifeste de façon de plus en plus marquée quand l'âge augmente.

Tableau 3: *Matrices des confusions: répartition des choix erronés sur chaque type d'intention par rapport au total des erreurs pour chaque groupe d'âge. Le total des % d'un tableau est égal à 100. Tableau a) résultats à 6 ans; b) à 8 ans; c) à 10 ans; d) à l'âge adulte.*

(a)

	Intention présentée				
	I1	**I2**	**I3**	**I4**	**T**
choix					
I1		14	10	12.3	36.3
I2	4.8		5	5	14.8
I3	8	11		6.5	25.5
I4	8	6.7	7.8		22.5

(b)

	Intention présentée				
	I1	**I2**	**I3**	**I4**	**T**
choix					
I1		9.9	12	17.2	38.9
I2	3.8		2.6	1.7	8.1
I3	7.6	12.9		8.2	28.7
I4	9.4	2.9	11.7		24

(c)

	Intention présentée				
	I1	**I2**	**I3**	**I4**	**T**
choix					
I1		6.8	14	22.4	33.2
I2	1.7		2.1	1.7	5.5
I3	3.4	7.6		11.9	30
I4	11.4	2.5	14.4		28.3

(d)

	Intention présentée				
	I1	**I2**	**I3**	**I4**	**T**
choix					
I1		0	16.1	20.3	36.4
I2	0.8		0	0	0.8
I3	11.9	0		17.8	29.7
I4	3.4	0.8	28.8		33

La comparaison des répartitions des pourcentages totaux en ligne montre que l'intention neutre sert en général de valeur-refuge et que la forme interrogative est de moins en moins choisie quand l'âge augmente. Cette configuration prosodique particulière de l'interrogation, produite aisément dès 7 ans, n'est donc pas encore détectée facilement à 6 ans ni même à 8 ans mais semble avoir un statut spécifique: elle pourrait exister sous forme de schéma permettant à tout le moins de conclure que si telle ou telle configuration prosodique entendue à un moment donné n'est pas identifiée, elle ne peut en tous cas correspondre à ce "schéma mental" de l'interrogation. L'analyse du corps des tableaux montre aussi deux "phénomènes stratégiques" intéressants:

- 1. les erreurs majoritaires (par rapport au total des confusions) à tout âge s'observent quand l'intention présentée exprime le regret (I4) et que les auditeurs choisissent l'intention neutre (I1). Les pourcentages de confusion entre I4 et I1 sont respectivement de 12.3 à 6 ans, 17.2 à 8 ans, 22.4 à 10 ans et 20.3 % pour les adultes. Cette confusion provient probablement du fait qu'aussi bien en terme de durée totale que de ton moyen, cette intention I4 est proche de I1, comme l'ont montré nos résultats lors des analyses des productions. Le choix erroné de I1 s'observe moins fréquemment lorsque l'intention exprimée est la joie, celle-ci étant plus différente de I1 en durée et fréquence moyenne, selon ces mêmes analyses de la production.

- 2. la confusion entre joie et regret augmente en valeur relative avec l'âge pour atteindre un fort pourcentage chez les adultes (respectivement 7.8 puis 11.7 puis 14.4 puis 28.8 % du total des erreurs quand l'âge augmente). Or I3 et I4 différaient plus nettement l'une de l'autre en durée et fréquence globales d'après les analyses de production. Ceci nous a semblé provenir d'une stratégie particulière liée à notre tâche (et non aux situations naturelles d'écoute), stratégie qui devient dominante avec l'âge: les auditeurs développeraient une tendance à se concentrer sur le dernier mot de chaque phrase et "ignoreraient" le début de la phrase. Cette stratégie est renforcée par le fait que les différences locales dues aux intentions sont également plus marquées en fin de phrase qu'au début comme l'ont montré les analyses de production.

CONCLUSION

En situation de **production** d'énoncés, par rapport à la forme affirmative, la forme interrogative s'est révélée être de durée plus courte et de contour montant aussi bien chez l'adulte que chez l'enfant. Les interrogatives se sont distinguées de toute autre forme expressive à la fois par le contour et la valeur moyenne de Fo (ce qui est connu depuis longtemps) mais aussi par la durée. La durée d'énonciation des autres intentions, et la fréquence fondamentale moyenne des mots qui composaient la phrase, se sont organisées en formes typiques, là encore aussi bien chez l'adulte que chez l'enfant: la joie était plus courte à exprimer que le regret, ce dernier se manifestant aussi par un ton moyen plus grave que la joie et des marges de variations de Fo plus restreintes. Adultes et enfants dès 7 ans partagent ces règles conversationnelles. Du côté de la **perception** des intentions, les processus de décodage semblent être plus tardivement installés. Même une forme de base comme l'interrogative reste difficile à identifier à 6 ou 8 ans lors du décodage du ruban sonore, tout en présidant sous forme schématique aux classements des énoncés comme "n'étant pas" du type interrogatif. Pour les autres intentions, les hiérarchies établies dans la production des intentions semblent en mesure d'expliquer les taux relatifs de confusion lors de la perception. Sous l'angle de la nature et de la fonction des représentations mentales qui gouvernent production et perception de la parole, ce type de données peut permettre de développer de nouvelles hypothèses.

DISCUSSION

La compréhension du langage ne relève pas seulement du décodage phonétique, de l'identification lexicale, des traitements syntaxiques et sémantiques, tous types de traitements obéissant chacun à sa propre loi de développement au cours de l'enfance, et dont la description occupe des pans entiers des études psycholinguistiques. La composante dite pragmatique qui s'y ajoute est beaucoup moins étudiée aussi bien chez l'adulte que chez l'enfant. La pragmatique qui forme un sous-domaine assez hétérogène du champ de la psycholinguistique, inclut des paramètres non linguistiques propres à la situation d'énonciation, où les intentions et émotions du locuteur prennent place (Moeschler, 1991; Jayez, 1989, Récanati, 1979, 1992), et ces déterminants des énoncés se marquent sur les caractéristiques intonatives et rythmiques du ruban sonore. Le langage est une activité coopérative qui, par conséquent, obéit aux mêmes lois que toute interaction sociale, comme Grice (1975) a été l'un des premiers à le rappeler. Des règles ou "maximes conversationnelles" constituent ces lois ou normes partagées qui gouvernent l'organisation des messages et qui peuvent être éventuellement délibérément violées pour tromper (mensonge) ou pour créer des effets spécifiques (ironie). Dans cette logique, la question se pose de séparer au cours du développement de l'enfant ce qui résulte d'une détermination quasi-biologique de

l'évolution de la voix selon tel ou tel affect ressenti (Scherer, 1984, 1986, Scherer & Zei, 1989), ou d'une simple imitation de la voix de l'entourage, et ce qui résulte de la prise de conscience de l'existence de ces règles et de leur utilisation grâce aux représentations mentales adéquates gouvernant tant la production que la perception de la parole. Il semble que la compréhension de la certitude soit réalisée dès 5 ans grâce à la configuration prosodique de l'énoncé (Moore, Harris & Patriquin, 1993). Mais les formes prosodiques de l'ironie ou du mensonge sont décodées plus tardivement: après 5 - 7 ans (Winner & Leekam, 1991). On peut supposer aussi que les formes prosodiques qui leur correspondent sont variées d'une part selon les locuteurs et instables d'autre part pour un même locuteur. Mais ceci demande à être vérifié. Les formes prosodiques expressives d'intentions plus fondamentales (joie, regret), qui peuvent être reliées à des déterminants biologiques de l'articulation au sens de Scherer (opus cit.), pourraient donner lieu à des configurations prosodiques plus contraignantes pour le locuteur, plus prototypiques pour l'auditeur. Il nous semble que le rythme d'énonciation et la distillation au cours du temps des marques intonatives jouent un grand rôle dans ces différenciations. Notre programme de recherche a pour ambition générale de valider ces hypothèses en poursuivant l'étude d'intentions diverses, tant en production qu'en perception. En y associant des études développementales, seules à même de trancher entre sophistication culturelle propre à une langue donnée et contraintes générales, nous espérons disposer d'un outil efficace de description des mouvements de la voix. Ce type de description intéresse aussi les chercheurs confrontés au problème de l'intégration des paramètres prosodiques dans les systèmes de synthèse et reconnaissance automatique de la parole (Cozannet, 1991; Pierrel & Carbonel, 1993).

Références:

Cozannet, A. (1991). Message to speech, Note interne, *Centre National d'Etudes des Télécommunications*.

Crystal, D. (1986). Prosodic development in language acquisition studies, in Fletcher, P. & Garma, M. (Eds.). *First language development*, (2d ed), Cambridge University Press, pp. 174-197.

Cutler, A. & Swinney, D.A. (1987). Prosody and the development of comprehension, *Journal of child language*, **14**, 147-167.

Gérard, C. & Rigaut, C. (1994). Patterns prosodiques et intentions des locuteurs: le rôle crucial des variables temporelles dans la parole, *Actes du 3ème Congrès Français d'Acoustique*.

Grice, H.P. (1975). Logic and conversation, in P. Cole and J.L. Morgan, Eds, *Syntax and Semantics*, vol. 3, Speech Acts. N.Y: Academic Press.

Jayez, J. (1989). Un aspect de l'analyse automatique des conversations: croyances, mouvements conversationnels, mouvements discursifs, *Cahiers de Linguistique Française*, **10**,147-170.

Moeschler, J. (1991). L'analyse pragmatique des conversations, *Cahiers de Linguistique Française*, **12**, 7-30.

Moore, C., Harris, L. & Patriquin, M. (1993). Lexical and prosodic cues in the comprehension of relative certainty, *Journal of child language*, **20**, 153-157.

Pierrel J.P. & Carbonel N. (1993). La reconnaissance automatique de la parole, *Le Courrier du CNRS*, **79**, 11.

Récanati, F. (1979). Insinuation et sous-entendu, *Communications*, **30**, 95-106.

Récanati, F. (1992). La pragmatique linguistique, *Le courrier du CNRS*, **79** p.21.

Scherer K.R. (1986). Vocal Affect Expression: a review and a model for future research, *Psychological Bulletin*, **99**, 143-165.

Scherer K. et Zei B. (1989). La voix comme indice affectif, *Revue médicale de la Suisse Romande*, **109**, 61-66.

Sorin, C. (1989). Perception de la parole continue, in M.C. Botte, G. Canevet, L. Demany, & C. Sorin, *Psychoacoustique et perception auditive*, série audition, INSERM/SFA/CNET, 123-139.

Winner, E. & Leekam, S. (1991). Distinguishing irony from deception: Understanding the speaker's second order intention. *British Journal of Developmental Psychology*, **9**, 257-270.

ETUDE DE L'INFLUENCE DES VARIATIONS PROSODIQUES DANS LES PROCESSUS DE TRAITEMENT DE LA PAROLE

DELPHINE DAHAN

UNIVERSITE RENE DESCARTES, PARIS
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE; EPHE

Résumé

Des recherches ont mis en évidence que la présence d'un accent d'emphase sur un mot facilite son traitement. Les manifestations acoustiques du mot accentué (allongement de sa durée, augmentation de sa fréquence fondamentale), ainsi que le contexte prosodique précédant, seraient responsables de cette facilitation. Le rôle de ce dernier serait de fournir des indices (sans doute rythmiques) qui permettraient à l'auditeur d'anticiper l'accentuation et de diriger ainsi son attention. Dans cette recherche, nous avons mesuré le temps de réaction à la détection du phonème initial du mot accentué, en fonction de son contexte (induisant ou non cet accent). Le temps de réaction à cette détection nous indique la vitesse du traitement de ce mot. Afin d'évaluer le rôle facilitateur du patron temporel précédant le mot accentué, nous avons utilisé les productions de deux locuteurs dont les manifestations temporelles différaient notablement lors de la production d'un accent d'emphase : le locuteur 2 accélèrait sa vitesse de parole avant la cible, alors que le locuteur 1 ne la modifiait pas. Ces deux locuteurs avaient par ailleurs des manifestations prosodiques similaires sur le mot accentué lui-même. Les résultats concernant les énoncés produits par le locuteur 2 indiquent qu'aucune variation prosodique ne semble avoir été véritablement utilisée par les auditeurs, même celles ayant lieu sur le mot-cible. Son débit de parole assez rapide pourrait expliquer ce résultat. En revanche, les résultats concernant les réalisations prosodiques du locuteur 1 suggèrent un faible rôle du contexte dans la facilitation du traitement du mot accentué. Bien que nos résultats ne puissent nous permettre de conclure aisément, il semble que les auditeurs n'utilisent que de façon modulée les variations prosodiques pour anticiper la configuration accentuelle à venir, en fonction des caractéristiques prosodiques des locuteurs.

Mots-clés: perception, prosodie, rythme de parole.

INTRODUCTION

Nos recherches portent sur la production et la perception des aspects prosodiques de la parole. Le phénomène stylistique particulier que nous étudions est l'accent d'emphase ou d'insistance. Cet accent occupe une place un peu à part dans l'organisation rythmique et accentuelle du français, car son rôle est essentiellement d'ordre sémantique et pragmatique (Garde, 1968; Rossi, 1979; 1985). En effet, l'accent d'emphase rend proéminent le mot sur lequel il porte, par rapport aux autres mots de l'énoncé. Cette proéminence focalise ainsi l'attention de l'auditeur sur ce mot. L'emphase signalerait l'importance de la proposition dans laquelle elle apparaît, la rendant essentielle à la signification du discours. Le processus de compréhension de l'énoncé serait donc modifié (Kimelman & MacNeil, 1987; Bean, Folkins, & Cooper, 1989). L'étude de la perception de ce phénomène peut nous permettre de comprendre plus généralement comment l'auditeur utilise les informations prosodiques du discours, afin d'accéder à sa signification.

L'ensemble des recherches portant sur la réalisation acoustique de cet accent a mis en évidence un allongement de la durée du mot accentué, ainsi qu'une notable augmentation de sa fréquence fondamentale (Séguinot, 1977; Cooper, Eady, & Mueller, 1985; Eady & Cooper, 1986). L'ensemble de ces manifestations rend le mot accentué saillant par rapport au reste de l'énoncé. Sa perception et son traitement, c'est-à-dire l'accès à son sens, sa fonction syntaxique, etc.., en sont ainsi facilités. Utilisant le paradigme de détection de phonème, de nombreuses recherches psycholinguistiques ont mis en évidence cette facilitation. En effet, le temps de réaction à la détection du phonème initial d'un mot est plus court lorsque ce mot est accentué que lorsqu'il ne l'est pas (Cutler & Foss, 1977).

Cutler (1976) a tenté de préciser, pour la langue anglaise, ce qui détermine la facilitation du traitement d'un mot accentué, en dissociant l'influence des variations prosodiques ayant lieu sur le mot accentué lui-même, de l'influence des variations apparaissant dans le contexte le précédant. En effet, les recherches de Martin et ses collaborateurs (Martin, 1972; Schields, McHugh, & Martin, 1974; Meltzer, Martin, Mills, Imhoff, & Zohar, 1976) ont suggéré que l'auditeur utilise de manière active le patron rythmique d'un énoncé, lui permettant d'effectuer des anticipations et focalisant ainsi son attention sur certains points de l'énoncé grâce au contexte prosodique précédant. Cutler a donc construit quatre versions d'une même phrase : dans une première version, le contexte prosodique induisait une focalisation sur le mot-cible de la phrase ; dans une seconde version, la focalisation était induite sur un autre mot de la phrase que le mot-cible ; enfin, dans les deux autres versions, Cutler a neutralisé les variations prosodiques de la cible (dans le premier cas, accentuée, dans le second cas, désaccentuée) en remplaçant ce mot par sa réalisation acoustique provenant d'une phrase neutre, sans focalisation. Les quatre versions étaient donc les suivantes (nous avons symbolisé l'accent d'emphase par la mise en majuscules du mot accentué; le mot-cible est souligné, c'est-à-dire *dirt* dans les quatre versions) :

(1) *She managed to remove the DIRT from the rug, but not the berry stains.*
(2) *She managed to remove the dirt from the RUG, but not from their clothes.*
(3) *She managed to remove the dirt from the rug, but not the berry stains.*
(4) *She managed to remove the dirt from the rug, but not from their clothes.*

Les temps de réaction à la détection du phonème initial du mot-cible (dans l'exemple, "d") étaient plus courts lorsque le contexte prosodique induisait sa focalisation (cas 1) que lorsque ce contexte induisait une focalisation sur un autre mot (cas 2). Cette facilitation a également été trouvée entre les cas 3 et 4, c'est-à-dire lorsque les réalisations acoustiques du mot-cible étaient identiques, mais l'ampleur de l'effet était sensiblement affaiblie. Ces résultats confirment que les auditeurs utilisent activement le contexte prosodique afin d'anticiper le signal à venir, et ainsi focalisent leur attention sur les focus induits par le contexte prosodique précédant. D'autre part, ces résultats soulignent l'importance de l'influence des manifestations prosodiques du mot accentué sur la facilitation de son traitement, contrairement à ce qu'avaient suggéré Shields et al. (1974).

Dans quelques recherches postérieures à celles-ci, Cutler et Darwin (1981) et Cutler (1987) ont cherché à préciser sur quels indices prosodiques du contexte les auditeurs se fondaient pour effectuer leurs anticipations. La courbe mélodique ne semble pas être un indice nécessaire à cette anticipation, puisque Cutler et Darwin (1981) ont retrouvé l'effet de facilitation en maintenant la fréquence fondamentale constante tout le long de la phrase. Cutler (1987) a créé artificiellement des variations de l'un des paramètres prosodiques (rythme, mélodie ou intensité) non congruentes avec les variations des deux autres, c'est-à-dire des variations induisant une focalisation sur un autre mot que les deux autres paramètres. Les résultats ont suggéré que, dans leur traitement des indices prosodiques, les auditeurs se fondent non pas sur une seule des trois dimensions prosodiques, mais exploitent leur cohérence.

L'ensemble de ces recherches nous semble souffrir de l'absence d'analyse acoustique des réalisations produites. Il est difficile de faire des hypothèses sur les indices potentiellement utilisés par les auditeurs sans savoir quels indices prégnants sont présents dans le signal. De plus, la manipulation des paramètres prosodiques opérée par Cutler (1987) a créé des réalisations sonores peu naturelles. Enfin, les phrases utilisées étaient prononcées par un seul locuteur. Nous pensons que les processus de perception et d'anticipation mis en place par les auditeurs peuvent être profondément influencés par les caractéristiques prosodiques des locuteurs. Cette recherche étudie le rôle du contexte prosodique dans le traitement d'une phrase. Nous avons choisi deux locuteurs masculins, présentant une configuration temporelle très contrastée (leurs manifestations prosodiques ont été présentées en détails dans Dahan & Bernard, 1994). En effet, lors de la production d'une emphase, le locuteur 1 ne modifie pratiquement pas sa vitesse de parole sur la portion de l'énoncé précédant le mot accentué (en comparaison avec le même énoncé produit sans insistance), alors que le locuteur 2 adopte un débit plus rapide avant et après le mot cible. La Figure 1 présente, pour chaque locuteur, les durées syllabiques, c'est-à-dire la durée moyenne d'une syllabe, mesurées sur la portion de phrase précédant le mot-cible, sur le mot-cible lui-même, et sur la portion de phrase suivant le mot-cible, dans la condition neutre et dans la condition avec un accent d'emphase.

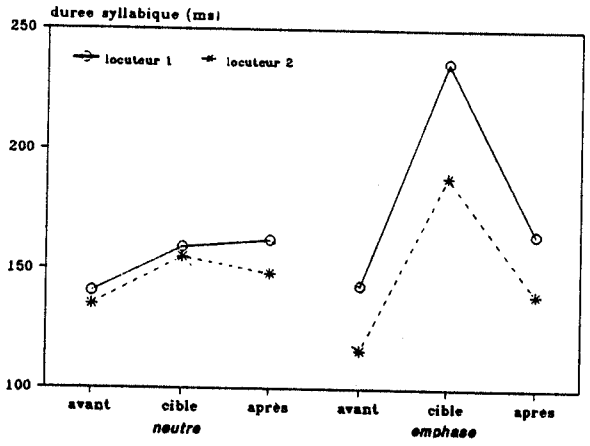


Figure 1 : durées syllabiques des locuteurs 1 et 2.

En condition neutre, c'est-à-dire sans intention d'insistance, le locuteur 2 a une vitesse de parole légèrement supérieure à celle du locuteur 1 (la durée syllabique est donc plus brève). En condition d'emphase, le locuteur 1 ralentit très fortement le mot accentué, mais sans changer sa vitesse de parole autour de la cible, alors que le locuteur 2, tout en produisant un ralentissement important sur la cible, présente une accélération très marquée avant la cible (la durée syllabique est alors beaucoup plus brève avec emphase que sans).

Par le choix de ces locuteurs, nous pouvons d'une part tenter de confirmer en français le rôle du contexte prosodique dans la facilitation du mot accentué, et d'autre part étudier l'utilisation de la configuration temporelle dans l'anticipation effectuée. Grâce aux rythmes de parole contrastés des deux locuteurs, nous avons pu manipuler le patron temporel de manière naturelle. Nous faisons l'hypothèse que les effets d'anticipation seront faibles ou de bien moindre ampleur lorsque la vitesse de parole avant le mot accentué est peu modifiée (locuteur 1), alors que l'effet sera de forte ampleur lorsque le locuteur fournit à l'auditeur des indices rythmiques importants pour anticiper l'insistance à venir et ainsi spontanément focaliser son attention sur le mot cible (locuteur 2).

METHODE

Matériel : chaque locuteur a été enregistré en chambre sourde. Il lui était demandé de lire les phrases présentées. Chaque phrase a été produite dans deux conditions : d'une part, le plus naturellement possible, sans volonté d'insistance particulière (intonation "neutre"), d'autre part en produisant une insistance sur un mot-cible, souligné et en caractères gras. Ces mots-cibles (un nom, un adjectif ou un adverbe) étaient choisis suffisamment éloignés du début et de la fin de l'énoncé. Nous avons sélectionné 16 phrases parmi un corpus plus étendu, phrases pour lesquelles une détection de phonème à l'initiale du mot-cible était possible (phonème consonnantique, par exemple).

A partir des deux versions de chaque phrase, que nous qualifierons respectivement de version "contexte neutre/cible neutre" et de version "contexte accentué/cible accentuée", nous avons construit deux nouvelles versions à l'aide d'un éditeur de signal sonore (Unice, par VECSYS). Nous avons placé le mot-cible accentué dans le contexte provenant de la version neutre, et nous avons placé le mot-cible neutre dans le contexte provenant de la version accentuée. Nous avons ainsi obtenu quatre versions de chaque phrase : (1) contexte neutre/cible neutre; (2) contexte accentué/cible accentuée; (3) contexte neutre/cible accentuée; (4) contexte accentué/cible neutre. Notre corpus était composé de 16 phrases X 4 versions X 2 locuteurs.

L'ensemble de ces phrases a été copié sur bandes de D.A.T. Huit bandes de 16 phrases ont été construites, chacune ne contenant qu'une seule des quatre versions de chacune des phrases. Quatre des huit bandes présentaient les phrases produites par le locuteur 1, les quatre autres, par le locuteur 2.

Procédure : Cent vingt auditeurs ont participé à cette expérience. Ils étaient dans la grande majorité des étudiants en Psychologie à l'Université Paris V René Descartes. Chaque auditeur était placé dans une pièce calme, et les phrases lui étaient présentées à l'aide d'écouteurs. Chaque auditeur n'entendait qu'une seule bande. Sa tâche était d'appuyer le plus rapidement possible sur le bouton-réponse dès qu'il percevait le phonème-cible à l'initiale d'un mot de la phrase. Ce phonème pouvait changer d'une phrase à l'autre et était donc précisé avant l'écoute de chaque phrase. Les temps de réaction à la détection du phonème, c'est-à-dire le délai temporel entre l'apparition du phonème et l'appui bouton, étaient mesurés. Nous avons également relevé les erreurs, c'est-à-dire les absences de détection.

RESULTATS

Les temps de réaction supérieurs à 1500 ms ont été exclus de l'analyse de variance, considérés comme témoignant d'un re-traitement tardif de la phrase par les auditeurs. Nous les avons tout de même intégrés dans le dénombrement des erreurs de détection, jugeant qu'ils témoignaient d'une plus grande difficulté à effectuer la tâche.

Les Figures 2 et 3 présentent respectivement les temps de réaction moyens et le nombre d'erreurs de détection, selon les locuteurs et les versions.

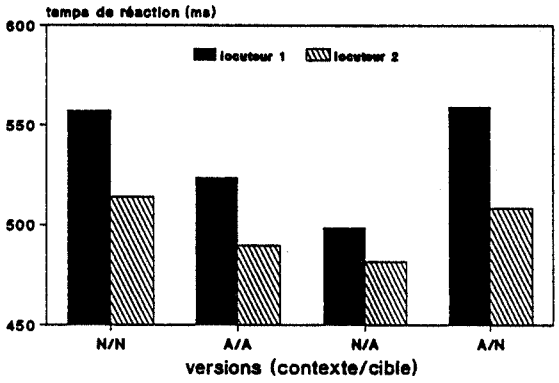


Figure 2 : temps de réaction

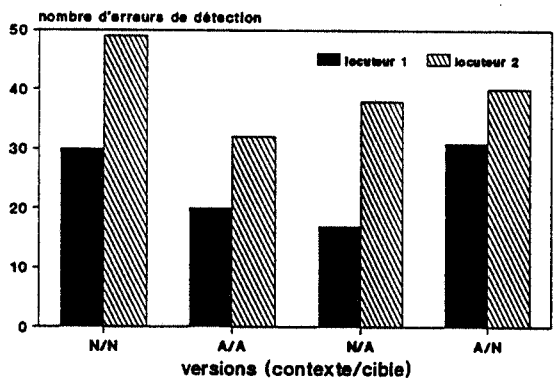


Figure 3 : nombre d'erreurs de détection.

Les temps de réaction sont plus longs pour les phrases prononcées par le locuteur 1 que ceux pour le locuteur 2. Ces temps présentent, pour le locuteur 1, une variation significative selon les versions auxquelles les auditeurs étaient confrontés (l'effet de la version est significatif, $F(3,177)=4.08$, $p<.01$) : lorsque la cible était accentuée, que le contexte induise ou non cet accent (A/A et N/A), les temps de réaction sont significativement plus courts que lorsque la cible était inaccentuée, que la contexte induise ou non un accent (N/N et A/N). La légère facilitation lorsque le contexte était neutre et la cible accentuée (N/A) par rapport à la version où contexte et cible étaient accentués (A/A), s'est avérée non significative. En revanche, les temps de réaction à la détection de phonème sur les énoncés du locuteur 2 ne présentent aucune variation significative ($F(3,177)=1.65$, NS), ni même l'effet facilitateur de l'accent d'emphase par rapport à la version neutre (A/A vs. N/N). L'analyse des erreurs de détection (Figure 3) permet de préciser ce résultat. En effet, les auditeurs ont effectué beaucoup plus d'erreurs de détection sur les phrases prononcées par le locuteur 2 que sur celles du locuteur 1. La vitesse de parole du premier, originellement plus rapide, et qui s'accélère lors de la production d'une insistance, a sans doute complexifié le traitement des énoncés, et donc la tâche de détection de phonème. Pour ce locuteur, les auditeurs semblent avoir souvent échoué à la détection du phonème-cible ; dans les cas où ils ont réussi cette détection, ils ne semblent pas avoir utilisé aucune des variations prosodiques contenues dans le signal de parole, ni celles sur le mot accentué, ni celles sur le contexte.

DISCUSSION

L'objectif de cette recherche était l'étude de l'effet du contexte prosodique sur le traitement du focus d'une phrase. En nous limitant aux résultats du locuteur 1, qui traduisent un traitement des informations prosodiques présentes dans le signal, nos résultats suggèrent l'absence d'un effet facilitateur du contexte. C'est en divergence avec les résultats de Cutler et ses collaborateurs sur la langue anglaise, présentés en introduction. Plusieurs explications peuvent être avancées. La première consiste à supposer que les différences dans l'organisation rythmique et accentuelle du français et de l'anglais seraient responsables de cette divergence. Cela nous paraît peu probable : de nombreuses expériences utilisant les méthodes de "tapping" ou de réitération de phrases en "ma-ma-ma" (voir en particulier Pasdeloup, 1990) ont montré que les auditeurs français parviennent aisément à extraire une structure rythmique fine des énoncés français. Une deuxième explication de ces résultats consisterait à penser que les auditeurs n'ont pas utilisé les

informations prosodiques mises à leur disposition pour anticiper l'accentuation à venir, dans le cadre de cette tâche très contraignante de détection de phonème. En effet, les auditeurs étaient informés de la possibilité pour le phonème-cible d'être à l'initiale de n'importe quel mot de la phrase. Les auditeurs auraient effectué leur traitement mot après mot, fragmentant ainsi perceptivement la continuité prosodique. Une consigne plus souple (indiquant la possibilité pour le phonème-cible de se trouver seulement à l'initiale de mots de contenu, par exemple), pourrait faire apparaître l'effet facilitateur du contexte. Enfin, l'explication que nous privilégierons, bien que l'absence de résultats pour le locuteur 2 ne puisse la confirmer fermement, est que le locuteur 1 ne fournit pas d'information rythmique suffisamment prégnante pour que les auditeurs s'en saisissent. Les auditeurs adapteraient leur stratégie de perception et d'anticipation à leur interlocuteur. Une nouvelle expérience, manipulant artificiellement la vitesse de parole de ce locuteur sur le contexte précédant (durées des segments, présence de pauses éventuelles, etc..), pourrait nous permettre de tester cette hypothèse, tout en gardant un rythme général suffisant pour permettre une détection aisée.

Néanmoins, la compréhension de l'utilisation des phénomènes prosodiques reste fondamentale, tant sur le plan théorique des modèles de perception de la parole continue, que pour son application à la reconnaissance automatique de la parole.

BIBLIOGRAPHIE

BEAN, C., FOLKINS, J.W., COOPER, W.E. (1989). The effects of emphasis on passage comprehension. *Journal of Speech and Hearing Research*, **32**, 707-712.

COOPER, W.E., EADY, S.J., MUELLER, P.R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, **77**, 2142-2156.

CUTLER, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics*, **20**, 55-60.

CUTLER, A. (1987). Components of prosodic effects in speech recognition. In *Proceedings of the XI International Congress on Phonetics Sciences*. Tallinn, Estonia, vol. 1, 84-87.

CUTLER, A., DARWIN, C.J. (1981). Phoneme-monitoring reaction time and preceding prosody: effects of stop closure duration and of fundamental frequency. *Perception and Psychophysics*, **29**, 217-224.

CUTLER, A., FOSS, D.J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, **20**, 1-10.

DAHAN, D., BERNARD, J.-M. (1994). Production et perception des phénomènes prosodiques dans la parole : comment identifie-t-on l'insistance malgré la variabilité inter-locuteurs ? *Actes du 3ème Congrès Français d'Acoustique*. Toulouse, mai 1994.

EADY, S.J., COOPER, W.E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, **80**, 402-415.

GARDE, P. (1968). *L'accent*. Paris: P.U.F.

KIMELMAN, M.D.Z., MacNEIL, M.R. (1987). An investigation of emphatic stress comprehension in adult aphasia: a replication. *Journal of Speech and Hearing Research*, **30**, 295-300.

MARTIN, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech. *Journal of Experimental Psychology*, **79**, 487-509.

MELTZER, R.H., MARTIN, J.G., MILLS, C.B., IMHOFF, D.L., ZOHAR, D. (1976). Reaction time to temporally-displaced phoneme targets in continuous speech. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 277-290.

PASELOUP, V. (1990). *Modèle de règles rythmiques du français appliqué à la synthèse de parole*. Thèse de Doctorat. Université de Provence, Aix-Marseille I.

ROSSI, M. (1979). Le français, langue sans accent ? In I. Fonagy, P.R. Léon (Eds). L'accent en français contemporain. *Studia Phonetica*, **15** (pp 13-51). Montréal: Didier.

ROSSI, M. (1985). L'intonation et l'organisation de l'énoncé. *Phonética*, **42**, 12-20.

SEGUINOT, A. (1977). L'accent d'insistance en français contemporain. In F. Carton, D. Hirst, A. Marschal, A. Séguinot (Eds). L'accent d'insistance. Emphatic stress. *Studia Phonetica*, **12** (pp 1-58). Montréal: Didier.

SHIELDS, J.L., McHUGH, A., MARTIN, J.G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, **102**, 250-255.

SYNTHESE DE PAROLE DE HAUTE QUALITE A PARTIR D'UN TEXTE : UNE COMPARAISON DE QUATRES ALGORITHMES CANDIDATS.

T. DUTOIT, H. LEICH

FACULTE POLYTECHNIQUE DE MONS,
31, BOULEVARD DOLEZ, B-7000 MONS, BELGIUM.

Résumé

Nous nous intéressons ici à l'utilisation de quatre modèles de parole dans le cadre de la conception d'un système de synthèse de parole de haute qualité à partir d'un texte (TTS : Text-To-Speech) par concaténation de segments? Nous examinons les problèmes rencontrés lors de l'élaboration de leurs modules de concaténation et d'adaptation de la prosodie et en comparons les performances en ce qui concernc : leurs possibilités de compression de la base de données, la charge de calcul associée à leurs algorithmes de synthèse, ainsi que leur intelligibilité et leur naturel.

Mots-clés: *synthèse, concaténation, adaptation de la prosodie, LPC, MBE, TD-PSOLA, MBR-PSOLA.*

INTRODUCTION

L'intelligibilité et le naturel offerts par un système TTS de haute qualité, pour lequel une fréquence d'échantillonnage de 16 kHz (ou plus) est généralement adoptée, dépendent de nombreux facteurs :

1. Le type de segments choisis.
2. Le corpus dont ils ont été extraits.
3. La qualité de la segmentation de ce corpus.
4. Le modèle de parole utilisé, qui conditionne en grande partie les algorithmes d'analyse et de synthèse.
5. L'importance de la dégradation apportée par l'étape de codage de la base de données des segments
6. L'efficacité de l'adaptation de la prosodie, qui dépend elle aussi fortement du modèle mis en oeuvre.
7. Les possibilité de lissage spectral offertes par l'algorithme de concaténation des segments.

Dans [8], nous avons examiné l'utilisation de quatre modèles susceptibles d'apporter la haute qualité recherchée, sur base d'une implantation logicielle des systèmes de synthèse TTS correspondants. Lors de leur conception, une seule et même base de données de segments (c-à-d la même segmentation du corpus enregistré) a été utilisée, de façon a mettre en évidence les contributions spécifiques des modèles sous-jacents.

Les modèles mis à l'examen étaient les suivants :

1. Le désormais classique modèle Auto-Régressif (LPC) [1], pour lequel un ordre de prédiction de 18 a été choisi, en accord avec la fréquence d'échantillonnage. Il a servi de base de comparaison.

2. Le modèle hybride Harmonique/Stochastique (H/S) proposé dans [2] (où il est appelé *modèle MBE*) et [3], qui décrit les signaux de parole comme une somme d'harmoniques dont les paramètres (amplitudes, fréquences, et phases) évoluent lentement, et de composantes de bruit. Sa caractéristique fondamentale est de transférer la décision Voisé/NonVoisé sur des bandes de fréquences, voire de la transformer en taux de voisement en fonction de la fréquence. Les degrés de liberté ainsi obtenus sont mis à profit pour une meilleure simulation des sons mixtes, pour lesquels bruit fricatif et vibrations périodiques peuvent coexister.

3. Le modèle 'nul', tel qu'implanté par l'algorithme de Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) [5], qui suscite un intérêt considérable étant donné son efficacité segmentale et supra-segmentale exceptionnelle, associée à une simplicité pratiquement inégalée.

4. Le modèle Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add (MBR-PSOLA), basé sur une re-synthèse H/S originale et efficace de la partie voisée des segments, avec une fréquence fondamentale et des phases initiales constantes [6].

Leurs particularités dans le contexte de la synthèse TTS de haute qualité sont développées dans la section suivante, qui est elle-même suivie d'une présentation des résultats de tests de qualité et d'intelligibilité menés en laboratoire. Un résumé comparatif vient en conclusion.

QUATRE SYNTHETISEURS

1. Le modèle Auto-Régressif.

Nous avons mis au point une synthèse LPC classique, avec un ordre de prédiction de 18. L'adaptation de la prosodie était immédiate (pitch et durées sont des paramètres explicites du modèle). Nous avons comparé les performances de concaténation des coefficients PARCOR et des LSP et constaté que, malgré l'existence d'un léger avantage théorique en faveur des LSP (Fig. 1), il est en pratique très difficile de distinguer les synthèses PARCOR et LSP d'une même

phrase. Pour ce qui est de la charge de calcul, nous avons implanté un filtre en treillis, ce qui correspond à 70 opérations par échantillons. Les coefficients du filtre étaient interpolés toutes les 5 ms. Quant aux possibilités de compression, des débits de 4000 bps peuvent être facilement obtenus, tout en restant transparents pour l'utilisateur.

2. Le modèle hybride H/S.

Une étude approfondie des erreurs d'analyse hybride H/S se trouve détaillée dans [4], où il apparaît que la précision de l'analyse est très sensible à une variation temporelle (même très faible) du pitch, ainsi qu'à une modification (plus importante) des amplitudes des sinusoïdes, et ceci quel que soit le critère d'analyse utilisé. Il y est montré que l'erreur qui en résulte apparaît principalement en haute fréquence, et qu'elle se retrouve à la synthèse sous la forme d'un bruit HF additif. En conséquence, nous avons veillé à ce que le corpus qui a servi de base à nos synthétiseurs soit enregistré avec un pitch pratiquement constant.

Un soin tout particulier a été apporté à la conception de l'algorithme de synthèse, pour lequel nous nous sommes volontairement écartés des techniques exposées dans [2], [3], ou même [11] (qui rapporte, à notre connaissance, la seule autre implantation du modèle H/S dans un système TTS). Nous leur avons préféré une méthode originale et plus rapide, qui calcule les échantillons de synthèse par recouvrement-addition (OLA) de transformées de Fourier inverse (IFFT) de fenêtres spectrales. Celles-ci sont obtenues par sommation de composantes stochastiques (sous la forme de bandes fréquencielles d'amplitude fixe et de phases aléatoires) et harmoniques (sous la forme des échantillons les plus significatifs de leurs noyaux de Dirichelet). La charge de calcul résultante a pu ainsi être réduite à environ 100 opérations par échantillon (c-à-d à peine 50% de plus qu'avec une filtre en treillis).

L'étape d'adaptation de la prosodie est légèrement plus complexe qu'en LPC. La durée et le pitch de chaque fenêtre d'OLA peuvent être facilement modifiés, mais il faut veiller à la continuité des phases, en propageant sur toutes les fenêtres de droite les changements de phases dus a une modification de la prosodie de la fenêtre courante.

En ce qui concerne la concaténation des segments, l'utilisation explicite, par le modèle H/S, de paramètres spectraux conduit tout naturellement à un lissage linéaire dans le domaine fréquenciel, qui est équivalent à une effet de fade in /fade out entre segments successifs (Fig. 1). Bien que théoriquement moins efficace que les transitions formantiques plus réalistes produites par l'interpolation de coefficients PARCOR, ce procédé conduit à des transitions très naturelles, a condition que les signaux de départ soient suffisamment proches au niveau acoustique (ce qui est pratiquement toujours le cas avec des diphones). De fait, nous avons vérifié qu'il devient pratiquement impossible de détecter les points de concaténation.

Quant au délicat problème de la discontinuité des phases au droit des frontières de segments, nous l'avons surmonté en ajoutant un offset à toutes les phases des segments à droite d'une frontière donnée (l'offset étant choisi de façon à gommer la discontinuité qui y est relative). Il en résulte une parole très fluide (les différences spectrales sont lissées), mais non exempte de bourdonnement (surtout dans les fricatives voisées et dans les semi-voyelles), qui est dû, selon nous, au non respect d'une certaine cohérence de phase à long terme, présente en parole naturelle, et irrémédiablement perdue lors de la segmentation du corpus.

Pour terminer, des débits de stockage de l'ordre de 10000 bps peuvent raisonnablement être obtenu avec le modèle H/S (qui a précisément été mis au point pour ses bonnes propriétés de codage).

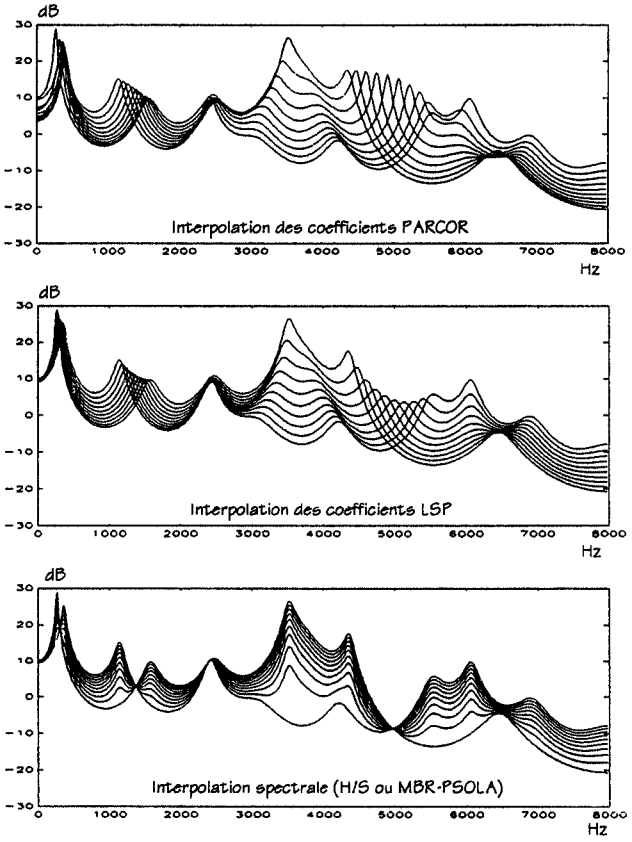


Figure 1. Interpolation linéaire des coefficients PARCOR, LSP, et hybrides H/S (et donc également des fenêtres MBR-PSOLA !) entre deux phones [l],au droit de la concaténation des diphones [al] et [lo].

3. L'algorithme TD-PSOLA.

Notre implantation du système de synthèse TTS basé sur TD-PSOLA est conforme à celle donnée dans

[5]. Ses inconvénients ont été soulignés dans [7] : un marquage de pitch optimal est difficilement compatible avec une préparation automatique de la base de données de segments (il a été réalisé manuellement dans notre cas), et des incompatibilités de phase, de pitch, et d'amplitude spectrale font ressortir les points de concaténation. Qui plus est, TD-PSOLA offre peu de possibilités de compression des segments (un débit de stockage de 80000 bps peut tout au plus être obtenu à l'aide d'un codage DPCM d'ordre 1). Des taux des compression plus importants nécessitent en effet la mise en oeuvre de codeurs dont la complexité dépasse largement celle du synthétiseur TD-PSOLA.

Par contre, TD-PSOLA conduit à une très bonne qualité segmentale (due au fait que cet algorithme ne fait implicitement usage d'aucun modèle de parole, et ne commet par conséquent aucune erreur de modélisation), et sa charge de calcul est remarquablement faible : à peine 7 opérations par échantillon.

4. Le modèle MBR-PSOLA.

Nous avons montré par ailleurs ([6], [7]) qu'une re-synthèse harmonique de la partie voisée des segments de synthèse, telle qu'effectuée par l'algorithme MBR-PSOLA, permet précisément d'éviter tous les inconvénients de TD-PSOLA, tout en continuant de bénéficier de son extraordinaire simplicité, et au prix d'une légère diminution de la qualité segmentale. Ceci résulte directement du fait que toutes les périodes voisées de la base de données de segments initiale se retrouvent stockés, dans la base MBR-PSOLA, avec un pitch unique et des phases initiales fixes. Les incohérences de pitch sont donc rendues inexistantes, l'opération de marquage de pitch n'est plus nécessaire, et une simple technique de lissage linéaire sur les fenêtres d'OLA (c-à-d au droit des échantillons) devient parfaitement équivalente à un lissage de leurs spectres d'amplitude (tel que celui réalisé en synthèse hybride H/S).

De plus, cette opération n'affecte en rien la simplicité de l'algorithme de synthèse initial, puisque la charge de calcul est maintenue à 7 opérations par échantillon (en incluant le lissage temporel linéaire)

En conséquence, MBR-PSOLA assure un maximum de fluidité à la parole synthétique. En contrepartie, un léger bourdonnement apparaît ça et là (vu cette fois le passage par un modèle), comme en hybride H/S.

Le dernier intérêt de MBR-PSOLA, et non des moindres, réside dans ses importantes possibilités de compression. En effet, l'opération de re-synthèse à pitch et phases initiales constantes assure une similarité maximale entre périodes successives, de sorte que la technique DPCM mentionnée plus haut est maintenant applicable entre échantillons homologues dans des périodes adjacentes (plutôt qu'entre échantillons successifs). Comme les formes d'onde évoluent lentement avec le temps (étant donné l'inertie propre au conduit vocal), et étant donné que les zones voisées couvrent à peu près 75 % de la base de données des segments (du moins pour le français, et dans le cas où les segments sont des diphones), on peut s'attendre à une réduction importante du débit de stockage, tout en maintenant la simplicité de la technique de codage initiale. Des tests sont en cours à ce sujet, et les résultats en seront dévoilés lors de la présentation.

TESTS DE QUALITE.

En nous conformant à l'approche du TNO [9], nous avons opté pour un combinaison de tests CVC et MOS pour estimer respectivement l'intelligibilité et le naturel de nos quatre synthétiseurs.

La méthodologie que nous avons suivie est basée sur [10]. Nous avons créé des listes phonétiquement équilibrées de 50 logatomes CVC, desquelles les semi-consonnes et les diphones inexistants ont été volontairement exclus. Les mots CVC ont été synthétisés avec un patron prosodique fixe, dans lequel le pitch était maintenu constant (110 Hz) et les durées imposées comme suit : [a, E, 9, i, O, y, u] = 70 ms, [e, 2, o, â, õ, ê] = 170 ms, fricatives = 100 ms, liquides = 80 ms, nasales = 80 ms, plosives = 100 ms. L'écoute de chaque mot était commandée directement par les auditeurs, en enfonçant une touche sur un clavier (et sans répétition possible). Chacun des 17 auditeurs a eu l'occasion de transcrire 4 listes CVC (une par synthétiseur), présentées dans un ordre aléatoire. Les stimuli ont été présentés au casque, dans une cabine insonorisée.

Des tests MOS ont également été menés pour estimer le naturel des synthétiseurs, ainsi que, plus spécifiquement, l'efficacité de leur technique de concaténation. Six longues phrases ont été synthétisées par chaque système, en imposant une prosodie obtenue par copie et stylisation d'une prosodie naturelle. Celle-ci était passée aux synthétiseurs sous la forme d'une suite de phonèmes avec leur durée et les positions et valeurs d'un ensemble de points de référence sur la courbe de pitch, supposée linéaire par morceaux. Les phénomènes micro-prosodiques n'ont pas été pris en compte. Les 24 stimuli résultants ont été présentés dans un ordre aléatoire aux auditeurs, à qui on demandait d'estimer séparément le naturel et la fluidité de la parole de synthèse, cette dernière appréciation étant fortement liée à la quantité de points de jonction laissés apparents par les algorithmes de concaténation. Les résultats de ces tests sont présentés à la Table 1.

Comme on pouvait s'y attendre, les tests CVC mettent clairement en évidence la supériorité des techniques hybride H/S, TD-PSOLA, et MBR-PSOLA sur la synthèse LPC. Un examen plus approfondi des résultats a par ailleurs montré qu'une confusion sur le

trait de voisement était à l'origine de la plupart des erreurs, principalement dans le cas des plosives. Des erreurs d'estimation de l'ouverture vocalique ([eE],[ao], ...) et du trait de nasalité ([âo],[aĩ], ...) ont également été rencontrées.

	LPC	hybride H/S	TD-PSOLA	MBR-PSOLA
intelligibilité CVC	54.6 %	65.7 %	78 %	72.8 %
fluidité MOS	50.4 %	73.5 %	65 %	75.6 %
naturel MOS	44.5 %	65 %	68.3 %	68.3 %

Table 1. Résultats des tests CVC et MOS.

Parmi les systèmes les plus intelligibles, TD-PSOLA maintient un léger avantage sur MBR-PSOLA. Comme annoncé précédemment, il est fort probable que ceci soit dû au fait que TD-PSOLA est intrinsèquement moins sensible aux erreurs d'estimation Voisé/Non Voisé que MBR-PSOLA. De fait, l'un comme l'autre étaient soumis à ce type d'erreur (dans le sens où le même algorithme de calcul et pitch a été utilisé dans les deux cas). Cependant, lorsqu'une fenêtre d'OLA voisée est considérée comme non voisée, ou vice-versa, TD-PSOLA se contente de lui appliquer un mauvais shift temporel par rapport à la précédente. Le voisement en lui-même ne s'en trouve pas vraiment affecté. Par contre, un fenêtre non-voisée considérée par MBR-PSOLA comme voisée est irrémédiablement re-synthétisée sous la forme d'une somme d'harmoniques.

Finalement, le synthétiseur MBR-PSOLA est lui-même légèrement mieux compris que son homologue hybride H/S, ce qui est un conséquence naturelle du fait que les fenêtres non-voisées sont laissées intactes par le re-synthèse MBR-PSOLA (de sorte qu'il n'y a d'ailleurs pas de différence entre TD et MBR-PSOLA pour ce qui les concerne), alors qu'elles sont soumises à synthèse (et donc à erreur de modélisation) en hybride H/S.

Quant au naturel, on aurait pu s'attendre à ce que les résultats MOS suivent la même tendance que les CVC : plus un synthétiseur fait usage d'un modèle de parole, moins il paraît naturel. Cependant, les systèmes hybride H/S et MBR-PSOLA l'emportent clairement sur leurs homologues TD-PSOLA et LPC en ce qui concerne la fluidité perçue, étant donné leurs possibilités de concaténation accrues. De sorte que, globalement, TD-PSOLA et MBR-PSOLA sont perçus comme également naturels, suivis de près par le modèle H/S, et loin devant le modèle AutoRegressif.

CONCLUSIONS.

Nos résultats sont résumés à la Table 2, qui peut se lire comme suit :

1. Les synthèses LPC, hybride H/S, et MBR-PSOLA sont supérieures à TD-PSOLA en ce qui concerne la disponibilité de procédures d'analyse (et donc de constitution de la base de donnée de synthèse) automatiques. Ceci est un point important lorsqu'on s'attache à mettre au point des systèmes TTS multi-lingues.
2. L'adaptation de la prosodie donne des résultats comparables pour les quatre systèmes
3. Pour ce qui est des possibilités de concaténation de segments, qui constituent un critère essentiel en synthèse TTS, la synthèse LPC est théoriquement légèrement meilleure que l'hybride H/S, qui est elle-même pratiquement équivalente à MBR-PSOLA. TD-PSOLA ne présente pratiquement aucune possibilité de concaténation.

Passant maintenant à des critères de nature plus économique, on peut dire que :

4. La disponibilité d'un algorithme de compression efficace est assurée pour les systèmes LPC et hybride H/S. Des recherches sont en cours pour MBR-PSOLA, mais il est clair que le taux de compression résultant sera supérieur à celui obtenu pour TD-PSOLA, tout en restant extrêmement simple.
5. En raison de la complexité de leurs algorithmes de synthèse respectifs, les approches LPC et hybride H/S nécessitent l'usage d'un DSP. Au contraire, tant TD que MBR-PSOLA tournent en temps réel sur un PC-386.

Finalement, lorsqu'on compare les résultats des tests d'intelligibilité et de naturel sur les quatre modèles, il apparaît que :

6. TD-PSOLA et MBR-PSOLA ont les meilleurs scores CVC, avec un léger avantage pour TD-PSOLA. Comme attendu, le synthétiseur hybride H/S est lui-même beaucoup plus intelligible que son homologue LPC.
7. La fluidité est meilleure en MBR-PSOLA et en hybride H/S, étant donné leurs possibilités de concaténation accrues.
8. En ce qui concerne le naturel, TD-PSOLA s'impose *de facto*, puisqu'il n'a recours à aucun modèle de parole. MBR-PSOLA, cependant, se hisse à sa hauteur, vu sa meilleure fluidité. H/S suit, loin devant la synthèse LPC.

Nous en concluons que MBR-PSOLA constitue une alternative intéressante à TD-PSOLA, surtout dans le contexte de la synthèse multi-lingue, pour laquelle la

possibilité de créer automatiquement des bases de données de segments, de les stocker sur une espace mémoire réduit, et d'en dégager une parole de haute qualité avec un nombre minimal d'opérations par échantillon, est d'un intérêt certain.

REFERENCES

[1] J.D. MARKEL, A.H. GRAY Jr, Linear Prediction of Speech, Springer Verlag, New York, pp. 10-42, 1976.

[2] D.W. GRIFFIN, J.S. LIM, 'Multi-Band Excitation Vocoder', IEEE Trans. on ASSP, vol. ASSP-36, pp. 1223-1235, august 1988.

[3] A.J. ABRANTES, J.S. MARQUES, I.M. TRANSCOSO, "Hybrid Sinusoïdal Modeling of Speech without Voicing Decision", EUROSPEECH 91, pp. 231-234.

[4] T. DUTOIT, H. LEICH, "An analysis of the performances of the MBE model when used in the context of a Text-To-Speech system", *Proc. EUROSPEECH 93*, Berlin, September 93, pp. 531-534.

[5] E. MOULINES, F. CHARPENTIER, "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", *Speech Communication*, Vol. 9, n°5-6. 1989.

[6] T. DUTOIT, H. LEICH, "Improving the TD-PSOLA Text-To-Speech Synthesizer with a Specially Designed MBE Re-Synthesis of the Segments Database", *Proc. EUSIPCO 92*, 25-28 august 92, Brussels, pp. 343-347.

[7] T. DUTOIT, H. LEICH, "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, Elsevier Publisher, december 1993.

[8] T. DUTOIT, *High Quality Text-To-Speech Synthesis of the French Language*, Ph.D. dissertation, october 1993.

[9] H.J.M. STEENEKEN, E. AGTERHUIS, *Speech intelligibility ans speech quality of seven speech coders for the HERMES project*, report IZF 1992 C-41, TNO Institute for Perception.

[10] H.J.M. STEENEKEN, *On measuring and predicting speech intelligibility*, Ph.D. dissertation, 1992, TNO Institute for Perception.

[11] E.R. BANGA, E. LOPEZ-GONZALO, C. GARCIA-MATEO, "A text-to-speech system for Spanish with a frequency domain based prosodic modification algorithm", *Proc. ICASSP 93*, Vol. 2, pp. 183-186.

	LPC	hybride H/S	TD-PSOLA	MBR-PSOLA
Analyse	automatique, facile	automatique, demande beaucoup de soin	semi-automatique (marquage de pitch)	automatique, demande beaucoup de soin
Codage (débit de stockage, taille de la base de données pour 3 min. à 16 kHz)	≈4 kBits/s ≈100 kbytes	≈10 kBits/s ≈200 kbytes	≈80 kBits/s ≈1.7 Mbytes	Technique de codage ad hoc en cours de test; meilleur que TD-PSOLA de toute façon
Adaptation de la prosodie	immédiate	simple, bien que moins immédiate	PSOLA lui-même	PSOLA lui-même
Concatenation des segments	lissage linéaire des PARCOR ou LSP ≈ transitions naturelles des formants	lissage linéaire ≡ fade in / fade out	faible, vu les incohérences de pitch, de phase, et d'enveloppe spectrale	cfr. hybride H/S pour les sons voisés et TD-PSOLA pour les autres
Synthèse	70 operations par échantillon	≈100 operations par échantillon, avec la méthode OLA/IFFT	7 operations par échantillon	5 operations par échantillon (7 avec le lissage)
Qualité de la modélisation	pauvre	très bonne	parfaite (pas de modèle)	très bonne, cfr. hybride H/S
Intelligibilité CVC	faible	élevée	presque parfaite	très élevée
Fluidité MOS	faible	très élevée	acceptable	very élevée
Naturel MOS	faible	élevée	très élevée	très élevée

Table 2. Comparaison des synthétiseurs LPC, hybride H/S, TD-PSOLA, et MBR-PSOLA.

RECHERCHE D'UNE STYLISATION EXTREME DES CONTOURS DE F0 EN VUE DE LEUR APPRENTISSAGE AUTOMATIQUE.

STEPHANIE DE TOURNEMIRE
Ecole Nationale Supérieure des Télécommunications - Paris

LAA/TSS/RCP
CNET-Lannion

Résumé

La synthèse de la parole à partir de texte utilise des informations d'ordre linguistique, pragmatique, prosodique et phonologique permettant de donner à la parole synthétique, une intonation naturelle. Etant donné un "modèle prosodique", permettant de rendre compte de tous ces niveaux d'information, on peut envisager d'utiliser des techniques d'apprentissage automatique pour extraire à partir des courbes mélodiques de parole naturelle, les contours mélodiques associés au modèle. Pour réaliser cet apprentissage, il est nécessaire de connaître les paramètres à extraire des courbes de F0 de la parole naturelle.

C'est dans ce contexte qu'un premier travail ayant pour objectif la recherche d'une stylisation extrême (c'est à dire maximale) a été réalisé. Cette stylisation maximale doit permettre de reproduire correctement la mélodie des phrases naturelles.

Ce travail a consisté en des tests perceptifs qui ont été réalisés sur différentes méthodes de stylisation de courbes mélodiques de phrases. Un premier objectif était d'analyser la pertinence prosodique de certaines unités phonologiques (voyelles, consonnes) et de différents points de ces unités (attaque, milieu, fin). Un second objectif était de vérifier l'hypothèse qu'une stylisation de la courbe mélodique à partir de 3 valeurs de F0 par mot était admissible. Les méthodes de stylisation sont décrites, l'expérimentation et les résultats des tests sont présentés et analysés.

Mots clés

Stylisation, Prosodie, Contours mélodiques, Apprentissage.

INTRODUCTION

La prosodie s'exprime du point de vue acoustique par 3 paramètres : l'énergie, la durée et l'évolution dans le temps de la fréquence fondamentale, F0.

La stylisation des contours mélodiques est basée sur l'hypothèse que la reproduction de tous les détails de la courbe de F0 n'est pas indispensable dans la mesure où ils ne sont pas tous perçus par l'oreille humaine. Elle consiste donc en une simplification de la courbe de F0. Cette simplification peut se faire suivant deux types de méthode correspondant à des approches théoriques différentes. Une première classe de méthodes que nous appellerons méthodes "supervisées" consiste à définir a priori, les points ou éléments de la courbe de F0 considérés comme pertinents et à interpoler entre ces points [3]. Une seconde classe de méthodes que nous appellerons méthodes "non supervisées" consiste à s'attacher à la forme globale de la courbe et à la simplifier indépendamment des unités prosodiques ([2], [4], [5]). Pour ces deux types de méthodes, l'interpolation entre les points cibles peut se faire à l'aide de droites ou de courbes (spline cubique, polynôme) [1]. Un outil de stylisation automatique permettant de mettre en oeuvre ces deux types de méthodes a été développé.

Dans le cadre d'une méthode de stylisation automatique supervisée, la recherche d'une stylisation maximale revient à étudier quelles sont les unités phonologiques ayant une valeur prosodique pertinente en contexte de phrase et au sein de ces unités, quelles sont les zones les plus représentatives de l'unité. Pour cette recherche, on utilisera donc des méthodes de stylisation "supervisées".

On s'intéressera plus particulièrement à la syllabe et au mot qui semblent être des unités prosodiques importantes [6]; en particulier en français où les accents mélodiques sont réalisés sur la première ou la dernière syllabe des mots. L'expérimentation permettra de

vérifier ou de rejeter l'hypothèse qu'une stylisation de F0 autour de valeurs de F0 sur les syllabes initiales et finales des mots serait suffisante pour styliser les phrases.

Les stylisations concernent uniquement les contours mélodiques. Tous les tests ont été réalisés en respectant les durées du signal naturel.

Nous présentons les tests perceptifs qui ont été menés afin de rechercher des stylisations maximales. La première partie de cet article décrit les différentes stylisations réalisées. La deuxième partie présente l'organisation des tests. Enfin, les résultats sont exposés dans la troisième partie.

1. LES METHODES DE STYLISATION

Les méthodes de stylisation testées ici, ont été choisies dans le but de chercher par approches successives, la stylisation maximale permettant de rendre compte de la mélodie des phrases. Traditionnellement, la littérature fait état de stylisations prosodiques (pour la synthèse de la parole) à l'aide de une, deux ou trois valeurs de F0 pour traduire la mélodie des phonèmes voisés [7]. Les stylisations présentées n'utiliseront donc jamais plus de 3 points par phonème.

Une première étape de segmentation manuelle des phrases en demi-phonèmes (marque au début, à la fin et sur le milieu des phonèmes) a été nécessaire pour pouvoir ensuite générer automatiquement les stylisations des courbes mélodiques de ces phrases. La valeur de F0 sur le milieu d'un phonème est extraite à la moitié de la durée totale du phonème. Le sonagramme ci-après donne un exemple de cette segmentation.

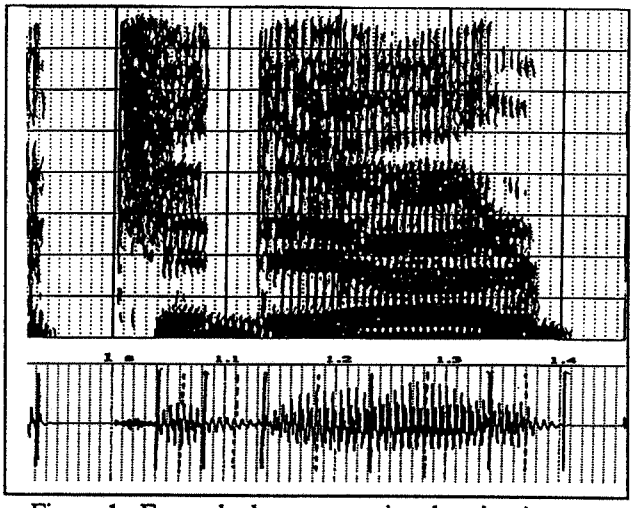


Figure 1 : Exemple de segmentation des phonèmes

Une fois cette segmentation réalisée pour toutes les phrases du corpus, les stylisations ont été générées automatiquement. Les contours mélodiques des différentes stylisations ont été appliqués sur les phrases naturelles par la technique PSOLA du système de synthèse du CNET [8].

Avant d'aborder la description des différentes méthodes de stylisation, il reste à préciser trois points méthodologiques :

Dans l'ensemble du document, chaque fois qu'il sera question de contours de phonème (voyelle ou consonne), il s'agira en fait du contour obtenu à partir de 3 valeurs de F0 (l'attaque, le milieu et la fin du phonème), interpolées entre elles linéairement.

De plus, il est très important de préciser que les stimuli de référence correspondent, eux aussi, à des stylisations de contours mélodiques. Ils sont choisis de manière à ce que la différence entre la stylisation testée et la stylisation de référence ne relève que des paramètres de la stylisation testée. Ainsi, tous les tests n'ont pas le même stimulus de référence et le contour mélodique originel (naturel) n'est jamais utilisé ici en stimulus de référence.

Enfin, l'interpolation entre les points est linéaire pour toutes les stylisations.

Les différentes méthodes de stylisation sont décrites ci-après. Pour chaque méthode, on précise l'objet de la stylisation c'est à dire les phénomènes que l'on cherche à étudier, la méthode de stylisation des phrases de test et de référence et la nature des phrases de test. Un exemple est donné pour chacune des stylisations sur une partie de la phrase "Mignonne, allons voir l'ombrageuse verdure ". Le contour naturel est indiqué par un trait fin et le contour stylisé par un trait épais.

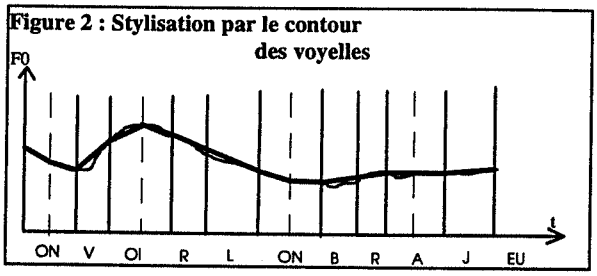
1.1. Stylisation par l'attaque ET le milieu ET la fin des voyelles (contours des voyelles)

Objet : Etudier si une interpolation linéaire de F0 de part et d'autre des voyelles (suppression de la microprosodie des consonnes) est suffisante pour rendre compte de la mélodie des phrases.

Points d'interpolation dans les phrases de test :

- L'attaque, le milieu ET la fin des voyelles,
- L'attaque du premier phonème du groupe de souffle si celui-ci est une consonne,
- La fin du dernier phonème du groupe de souffle si celui-ci est une consonne.

Dans la suite du document, cette méthode de stylisation sera notée CV (Contour Voyelles).



Stylisation de référence : Les points d'interpolation de la stylisation de référence sont donnés ci après :

- l'attaque, le milieu et la fin de toutes les voyelles,
- l'attaque, le milieu et la fin de toutes les consonnes voisées.

Dans la suite du document, cette méthode de stylisation sera notée CP (Contour Phonèmes).

Caractéristiques des phrases : Les phrases comportent tous les types de consonnes voisées (des occlusives sonores, des fricatives sonores et des nasales). Les consonnes interviennent dans différentes positions du mot (voir annexe : phrases n° 1, 2, 3, 4, 5).

1.2. Stylisation par l'attaque OU le milieu OU la fin des voyelles

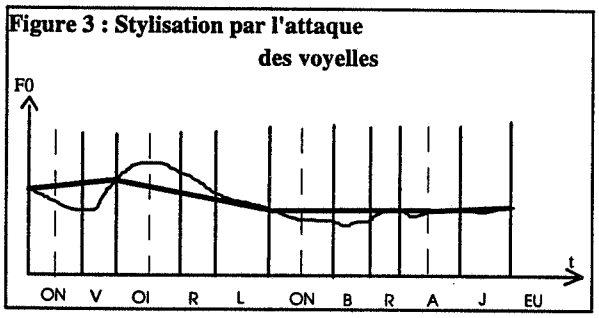
Objet : Etudier si l'une des trois valeurs de F0 (l'attaque, le milieu ou la fin des voyelles) peut suffire à styliser le contour de la voyelle [9]. Pour cette étude, trois stylisations ont été réalisées.

***Points d'interpolation* pour les phrases de test :**

- l'attaque (notée AV) OU le milieu (notée MV) OU la fin des voyelles (notée FV).

Et dans tous les cas :

- L'attaque du premier phonème du groupe de souffle si celui-ci est une consonne,
- La fin du dernier phonème du groupe de souffle si celui-ci est une consonne.



Stylisation de référence : CV

Caractéristiques des phrases : Les phrases comportent tous les types de voyelles (voyelles orales, nasales et semi-voyelles), des mots mono et plurisyllabiques, des phrases énonciatives et interrogatives. (Voir annexe : phrases n° 6, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23).

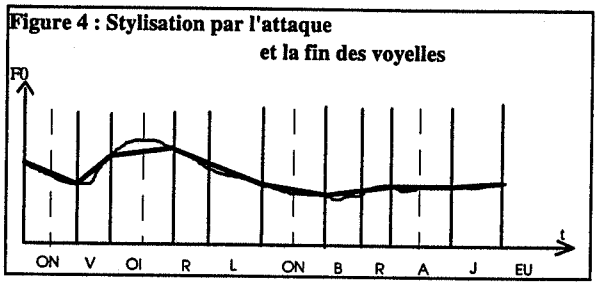
1.3. Stylisation par attaque et fin des voyelles

Objet : Etudier si les trois points : l'attaque, le milieu et la fin des voyelles sont nécessaires, ou si deux points (début et fin des voyelles) suffisent pour la stylisation.

Points d'interpolation pour les phrases du test :

- l'attaque et la fin des voyelles,
- L'attaque du premier phonème du groupe de souffle si celui-ci est une consonne,
- La fin du dernier phonème du groupe de souffle si celui-ci est une consonne.

Dans la suite du document, cette méthode de stylisation sera notée AFV (Attaque et Fin des Voyelles).



Stimulus de référence : CV

Caractéristiques des phrases : Les phrases comportent des voyelles susceptibles d'avoir des schémas en "hat pattern" (F0 montant-descendant) (Voir annexe : phrases n° 16, 24, 25, 26, 27, 28, 29).

1.4. Stylisation par mot

Objet : En français, l'accentuation est généralement réalisée sur la première et la dernière syllabe des mots. L'objet de ce test est donc d'étudier si une stylisation par mot ne tenant compte que de la première et de la dernière syllabe des mots pouvait être suffisante pour exprimer la mélodie de la phrase et quelles étaient ses limites.

Une approche possible était de prendre les valeurs de F0 au début et à la fin du contour mélodique du mot ainsi que sur son maximum intermédiaire éventuel. Cependant, les montées de F0 en milieu de mot sont généralement dues à des effets d'emphase ou de rythme. Dans le cadre de l'apprentissage de contours prosodiques en vue de leur génération automatique, on désire extraire des contours "neutres", sans emphase, que l'on pourrait appeler contours

"intrinsèques" des mots. Ces contours minimaux pourraient ensuite être enrichis par des effets d'emphase, de focus ou de rythme.

Nous avons donc choisi une stylisation alternative décrite ci-dessous :

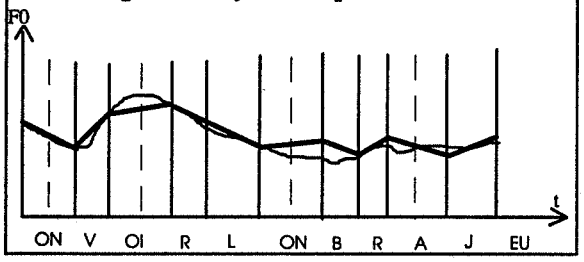
Points d'interpolation des phrases de test :

- L'attaque de la première voyelle des mots,
- L'attaque et la fin de la dernière voyelle des mots,
- L'attaque du premier phonème du groupe de souffle si celui-ci est une consonne,
- La fin du dernier phonème du groupe de souffle si celui ci est une consonne.

Dans la suite du document, cette méthode de stylisation sera notée MOT.

Quand un contour est linéaire, sur plus de deux phonèmes, une fluctuation (± 7 Hz) est appliquée au début et à la fin des phonèmes (consonnes et voyelles) telle que l'illustre la figure 5 pour les phonèmes "ON-B-R-A-J".

Figure 5 : Stylisation par mot



Stimulus de référence : CP

Caractéristiques des phrases : Phrases comportant des mots d'au moins 3 syllabes. (Voir annexe : phrases n° 3, 5, 13, 14, 16, 18, 19, 20, 21, 24, 25, 27, 31, 32, 33, 34).

2. DEROULEMENT DES TESTS PERCEPTIFS

Quatorze (14) auditeurs naïfs, non rémunérés, ont participé aux tests. Les stimuli étaient présentés en ordre aléatoire.

Quatre tests de dégradation (DCR : Dégradation Catégorie Rating [10]) et un test de préférence [11] ont été réalisés :

1. Test de dégradation pour la stylisation par le contour mélodique des voyelles (CV, Réf : CP)
2. Tests de dégradation pour la stylisation par
   a. l'attaque des voyelles (AV, Réf : CV)
   b. le milieu des voyelles (MV, Réf : CV)
   c. la fin des voyelles (FV, Réf : CV)
3. Test de dégradation pour la stylisation par l'attaque et la fin des voyelles (AFV, Réf : CV)
4. Test de dégradation pour la stylisation par mot (MOT, Réf : CP)
5. Test de préférence par paires pour la stylisation par mot (MOT/CP)

Les auditeurs ont écouté sur casque des paires de phrases stylisées séparées par un silence de 500 ms. Les auditeurs disposaient de 5 secondes pour enregistrer leur réponse. Chaque séance commençait par une courte séance d'apprentissage.

Pour les tests de dégradation, les auditeurs devaient noter de 1 à 5 (dégradation très gênante (1), dégradation gênante (2), dégradation légèrement gênante (3), dégradation perceptible mais pas gênante (4), dégradation imperceptible (5)).

Pour le test de préférence, les auditeurs devaient choisir entre 1. "Préfère la première", 2. "Préfère la deuxième, 3. "Pas de préférence". Les paires étaient présentées de façon contre balancée (A-B et B-A).

3. RESULTATS

Les résultats sont présentés ici essentiellement sous la forme de tableaux présentant les moyennes (et écart-types) des notes obtenues.

Il faut noter que pendant les tests de dégradation, des auditeurs ont fait remarquer qu'ils étaient gênés par l'appellation de "dégradation" et que ce terme pouvait, dans le cadre du test, être remplacée par "différence". En effet, il n'existe pas **une bonne** prosodie pour une phrase donnée mais plusieurs en fonction des locuteurs et du cadre pragmatique et sémantique de l'énoncé.

3.1. Tests de dégradation

3.1.1. Tests sur la stylisation par l'attaque ET le milieu ET la fin des voyelles (contour des voyelles)

Tableau 1 : Test de dégradation de CV /CP

Condition	Moyenne	Ecart-type
CV / CP	4.59	(0.63)

La valeur moyenne de 4.59 se trouve entre "dégradation imperceptible" et "dégradation perceptible non gênante". Cela signifie que les auditeurs ne sont pas gênés par l'absence de micromélodie sur les consonnes voisées : une interpolation du contour de F0 de part et d'autre des contours des voyelles est donc tout à fait admissible.

3.1.2. Test de la stylisation par l'attaque OU le milieu OU fin des voyelles

Tableau 2 : Test de dégradation de AV/CV, MV/CV et FV/CV

Condition	Moyenne	Ecart-type
AV / CV	4.06	(0.85)
MV / CV	4.42	(0.81)
FV / CV	3.97	(1.02)

Les notes moyennes sont situées entre "dégradation imperceptible" et "dégradation perceptible non gênante".

Un test t [12] a été calculé afin d'étudier si les différences de résultats entre les stylisations par l'attaque, le milieu et la fin des voyelles étaient significatives (14 auditeurs ont écouté 16 paires, donc on a N = 224 et $t_{224} = 2{,}60$).

Tableau 3 : Valeurs du test t entre AV, MV et FV

Comparaison	Test t
Test t entre AV et MV	4.69 (S)
Test t entre AV et FV	1.01 (NS)
Test t entre MV et FV	5.25 (S)

Les différences entre AV et MV et entre FV et MV sont significatives. La stylisation par le milieu des voyelles est donc préférée, de manière significative, aux stylisations par l'attaque ou par la fin des voyelles. Le milieu de la voyelle serait donc la partie la plus pertinente et la plus représentative du contour de la voyelle. Ce résultat s'explique facilement par le fait que la plupart des voyelles ont un contour plat, montant ou descendant, la valeur milieu est donc la valeur moyenne du contour.

3.1.3. Test sur la stylisation par l'attaque ET la fin des voyelles

Tableau 4 : Test de dégradation de AFV / CV

Condition	Moyenne	Ecart-type
AFV / CV	4.38	0.74

La valeur de 4.38 indique un jugement intermédiaire entre "dégradation imperceptible" et "dégradation perceptible non gênante". Ceci signifie que les auditeurs ne sont pas gênés par le non-respect des valeurs de F0 au centre des voyelles mais perçoivent une différence.

Une analyse plus précise de la répartition des notes fait apparaître 12 % de réponses "dégradation perceptible légèrement gênante". Elles correspondent à des phrases comportant des mots monosyllabiques (phrases 16 et 28) sur lesquels on trouve des schémas en forme de chapeau.

La prise en compte des valeurs de F0 initiales et finales des voyelles s'avère efficace dans la plupart des cas pour rendre compte de la mélodie des voyelles, mis à part certains cas d'emphase réalisés par des formes mélodiques de type "chapeau". Cependant, dans la recherche d'une stylisation extrême où l'objectif ne consiste pas à rechercher la réalisation de toutes les nuances prosodiques, une stylisation par l'attaque et la fin des voyelles serait envisageable. Nous allons voir qu'une réduction encore plus grande des mouvements mélodiques se révèle tout à fait acceptable.

3.1.3. Stylisation par mot

Tableau 5 : Test de dégradation de MOT / CP

Condition	Moyenne	Ecart-type
MOT / CP	4.17	0.79

La valeur de 4.17 indique que les auditeurs perçoivent dans la majorité des cas "une dégradation perceptible non gênante".

Une analyse plus précise de la répartition des notes fait apparaître 16 % de réponse "dégradation perceptible légèrement gênante". Elles correspondent à des mots longs tels que "monosyllabique" ou "anticonstitutionnel". En effet, la stylisation par mot ne permet pas de rendre compte des accents présents au milieu des mots.

Pendant le test, les auditeurs ont fait remarquer que la stylisation entraînait dans certains cas des différences de sens : les effets d'emphase exprimés au milieu des mots, n'apparaissent plus sur les phrases stylisées. Malgré cela, la stylisation était tout à fait acceptable.

Le test de préférence par paires, dont nous allons donner les résultats ci après, permet de compléter l'analyse de cette stylisation.

3.2. Test de préférence

Tableau 6 : Test de préférence entre CP et MOT

Condition	Préfère CP	Préfère MOT	Sans préf.
CP - MOT	42.6 %	40.2 %	17.2 %

Une analyse sur les données [13] a montré que la préférence pour la stylisation CP n'était pas significative.

Les résultats de ce dernier test confirment donc le fait qu'une stylisation par mot peut être envisagée pour l'extraction des contours mélodiques.

CONCLUSION

Des tests perceptifs sur des méthodes de stylisation ont été réalisés. Les résultats obtenus permettent de tirer les conclusions suivantes.

La présence ou l'absence de la micromélodie des consonnes voisées n'est pas perceptible par un auditeur naïf.

Les stylisations par l'attaque, le milieu ou la fin de la voyelle entraînent une dégradation perceptible non gênante, la stylisation par le milieu des voyelles étant la moins dégradée.

La stylisation par seulement l'attaque et la fin des voyelles par rapport à une stylisation par 3 points (attaque, milieu et fin) entraîne un jugement intermédiaire entre dégradation imperceptible à perceptible non gênante.

Enfin, la stylisation par 3 points par mot entraîne une dégradation perceptible non gênante. La comparaison entre une stylisation par mot et une stylisation par contour de phonème ne donne pas de préférence significative pour l'une ou pour l'autre.

La stylisation par mot qui permet de représenter la courbe mélodique par trois valeurs fixes de F0 du mot quelle qu'en soit la longueur semble donc tout à fait acceptable comme stylisation initiale. Ceci présente un intérêt important pour l'extraction et la génération automatique de contours prosodiques.

Pour réaliser l'apprentissage des contours mélodiques des phrases, nous pouvons donc choisir une stylisation à partir de trois points par mots (l'attaque de la première voyelle du mot et l'attaque et la fin de la dernière voyelle du mot). Cette stylisation correspond à la mélodie "intrinsèque" des mots dans les phrases. Nous savons qu'elle ne permet pas de rendre compte des effets d'emphase qui peuvent apparaître en milieu de mot. Néanmoins, à partir de cette mélodie "intrinsèque", il sera possible d'élaborer des règles de superposition de contours mélodiques, permettant de rendre de compte des effets d'emphase. Ces règles permettront aussi d'introduire de la variabilité [14] afin de rompre la monotonie de la mélodie pour la synthèse de la parole à partir du texte (TTS). Une analyse plus détaillée des résultats du test 3.1.4 pourra aider à modéliser ces règles.

Remerciements

Je tiens à remercier Martine Boyer du département LAA/TSS/CMC qui m'a aidée à la préparation et la réalisation des tests.

Références

[1] t'Hart J. "F0 stylisation in speech. Straight line versus parabolas" J. Acoust. Soc. Am. 90 (6), december 1991.

[2] Hirst D.J., Nicolas P., Espesser R. (1991) "Coding the F0 of a continuous text in french : an experimental approach." Proc. ICPhs 12 (Aix), 5, 234-237.

[3] Rossi M., Di Cristo A., Hirst D., Martin P., Nishinuma Y., "Procédure de stylisation et applications", dans "L'intonation, de l'acoustique à la sémantique" Klincksieck, pp. 30-82

[4] Beaugendre F., d'Alessandro C., Lacheret-Dujour A., Terken J., "A perceptual study of french intonation."

[5] Ten Bosch L., "On the automatic classification of pitch movements " IPO papers for Eurospeech 93 and ESCA Workshop on prosody. Sept. 1993.

[6] Vaissière J., "Variance and Invariance at the world level", dans Invariance and Variability in Speech Processes, édité par J.S. Perkel et D.H. Klatt, Lawrence Erbaum Associates, Publishers, 1986 pp. 534-539.

[7] La parole et son traitement automatique. Calliope. p 430

[8] Moulines E., Charpentier F., "Pitch-Synchronous waveform processing techniques for text-to-speech synthesis using diphones". Speech Communication, décembre 1990, Vol. 9, n° 5 et 6, pp. 453-467.

[9] Vaissière J. "The use of prosodic parameters in automatic speech recognition" Note Technique Cnet LAA/TSS/353, décembre 1987.

[10] Méthode d'évaluation DCR : Degradation Category Rating CCITT 1984

[11] La parole et son traitement automatique. Calliope p. 472-475.

[12] Experimental design in Psychological research. Allen L. Edwards. Holt Rinehart Winston.

[13] CCITT ad-hoc group of experts on 16 Kbits speech coding. WD 2/21, Mai 1992.

[14] Delais E. "Prédiction de la variabilité dans la distribution des accents et les découpages prosodiques en français". A paraître dans les JEP 94,

ANNEXE : CORPUS

1. Ce marin est brave
2. Cet arrosoir est idéal pour mon jardin
3. L'agglomération de Lyon est gigantesque
4. Le voleur de poires a été retrouvé
5. Partir loin pour longtemps, quel dépaysement !
6. J'en ai assez de vos histoires
7. Voulez-vous vraiment partir ?
8. Quel train prenez-vous ?
9. Vous prenez quel train ?
10. Vous voulez danser la valse.
11. Vous voulez danser la valse ?
12. Vous devez danser.
13. Vous confirmez la réservation.
14. Vous confirmez la réservation ?
15. Le ballon est tombé dans le feu
16. Luc adore les tourterelles
17. Philippe joue de la flûte
18. J'ai peur d'essayer
19. Cela paraît difficile
20. Ces proses sont de véritables merveilles
21. Les américains sont à Toulouse
22. Demain me paraît loin
23. Un beau chant m'a réveillé
24. Ce magnifique animal s'est échappé.
25. Je vais vous raconter une histoire extraordinaire
26. On se battait avec des couteaux dans le faubourg de la ville
27. Faire un referendum sans le peuple c'est anticonstitutionnel
28. Etes-vous prêt ?
29. Une corbeille pleine de fruits est posée sur la table
30. Le Guatemala est un pays magnifique
31. Ce mot est vraisemblablement monosyllabique
32. Je soulève un point excessivement délicat
33. Glisser sur le toboggan est l'activité préférée des enfants
34. Les libellules sont d'invraisemblables voyageuses
35. Mignonne allons voir l'ombrageuse verdure.

LA PERCEPTION DE L'ACCENT EN FRANÇAIS DU QUÉBEC ET EN FRANÇAIS DE FRANCE

DENISE DESHAIES†, ISABELLE GUAÏTELLA‡, CLAUDE PARADIS†

†C.I.R.A.L., UNIVERSITÉ LAVAL, QUÉBEC
‡C.N.R.S. URA 261, INSTITUT DE PHONÉTIQUE, AIX-EN-PROVENCE

Résumé

L'accent, notion qui est au coeur des travaux en phonétique et en phonologie, constitue toujours une problématique complexe, compte tenu notamment du manque d'études exhaustives du système prosodique de diverses variétés linguistiques utilisées en contexte spontané (Paradis 1993). C'est dans le cadre de cette problématique que nous avons mis au point une expérimentation qui visait à vérifier si des sujets francophones appartenant à deux communautés linguistiques différentes (Québec et France) et ayant une expertise différente en phonétique (phonéticiens et non phonéticiens) avaient recours aux mêmes stratégies dans la perception de l'organisation accentuelle de séquences extraites de corpus spontanés de français du Québec et de français de France et dont les premiers résultats seront présentés dans cet exposé.

Mots-clés: accent, variation, perception, français québécois, français de France

INTRODUCTION

Les études sur l'accentuation en anglais et en français ont mis en évidence le rôle des indices acoustiques dans la perception des syllabes accentuées. Cependant, l'importance accordée aux variations d'intensité, de durée segmentale et de fréquence fondamentale a été très variable suivant les auteurs et les périodes et il semble difficile d'identifier un paramètre ou un ensemble de paramètres qui soit indiscutablement associé à l'accent (Bruneau 1931; Delattre 1938; Bolinger 1958; Fry 1958; Lieberman 1960; Rigault 1962; Morton & Jassem 1965; Boudreault 1968; Di Cristo 1978, 1981, sous presse; Warren & Santerre 1979; Vaissière 1980; Hirst & Di Cristo 1984; Rossi *et al.*1981; Guaïtella 1991).
Cet état de fait peut être dû à plusieurs causes: soit nous ne savons pas différencier des accents de nature différente, soit il n'y a pas d'adéquation entre les indices acoustiques envisagés et les indices perceptuels, soit nous négligeons des indices pertinents appartenant à d'autres niveaux (grammatical, sémantique, pragmatique, etc.). Par ailleurs, on peut admettre que chaque langue dispose de spécificités dans l'utilisation des indices acoustiques et que les individus, ou les groupes d'individus, appliquent des stratégies diverses dans la production et la perception de l'accent.
Il existe un certain nombre d'études sur la perception de l'accent, ainsi que sur d'autres phénomènes suprasegmentaux, par des sujets linguistiquement entraînés et non entraînés. Ces études montrent que les taux de perception pour les deux catégories de sujets sont comparables dans une certaine mesure (Lieberman 1965; Allen 1972; McDowall 1974; Paradis & Deshaies 1990; Mertens 1991). Les travaux de Lieberman et de Mc Dowall montrent que les réponses sont satisfaisantes lorsqu'on demande aux sujets (naïfs ou non) de juger de façon binaire le caractère proéminent d'une syllabe (accentuée vs non accentuée), mais que les résultats sont beaucoup moins satisfaisants quand les sujets utilisent un système de notation de type continu (Dittman & Wynne 1961; Lieberman 1965; Allen 1972).

PROBLÉMATIQUE

Si nous savons que la perception d'un phénomène linguistique est largement déterminée par la langue maternelle de l'auditeur (Singh & Black 1966; Abramson & Lisker 1970; Paradis, Dolbec & Brousseau 1993) et par sa connaissance des facteurs phonétiquement pertinents, nous savons encore peu de choses sur les facteurs psychologiques et sociologiques qui interviennent dans la perception d'éléments linguistiques, même si, du moins pour la perception de l'accent, on croit que diverses stratégies sont susceptibles d'intervenir (Fónagy 1979; Paradis & Deshaies 1990; Guaïtella 1991).
C'est dans ce cadre que nous avons mis au point une expérimentation qui visait à vérifier si des sujets francophones appartenant à deux communautés linguistiques différentes (Québec et France) et ayant une expertise différente en phonétique (phonéticiens et non phoné-

ticiens) avaient recours aux mêmes stratégies dans la perception de l'organisation accentuelle de séquences extraites de corpus spontanés de français du Québec et de français de France.

HYPOTHÈSES

Les hypothèses générales qui sous-tendent cette recherche sont les suivantes:
· l'accent est une catégorie linguistique qui peut être théoriquement réalisée par un nombre infini de niveaux de proéminence. Nous pouvons donc supposer que plus une syllabe sera jugée comme accentuée par un nombre important de sujets, plus elle pourra être considérée comme proéminente (cf. Allen 1972);
· Dans la tâche d'identification des syllabes accentuées en parole spontanée, les sujets des deux communautés devraient être plus ou moins sensibles aux mêmes indices acoustiques pour les syllabes les plus proéminentes. Bien que ces indices (variations de F_0, intensité et durée segmentale) soient bien connus, l'importance des interactions de ces indices entre eux ainsi qu'avec d'autres paramètres est par ailleurs beaucoup moins claire qu'il n'y paraît au premier abord;
· Pour les syllabes dont la proéminence est moindre, nous supposons que les sujets auront recours à des stratégies différentes de traitement des indices acoustiques. Ces stratégies pourront être de type idiosyncratique ou être liées à l'appartenance à une communauté géo-socio-linguistique particulière ou au degré d'expertise en phonétique.

MÉTHODOLOGIE

A- TESTS DE PERCEPTION

• *Test «Québec»* (TestQ)
20 séquences sonores ont été extraites d'une entrevue sociolinguistique d'un homme de 32 ans de classe moyenne et habitant dans la région de Chicoutimi-Jonquière (Québec).

• *Test «France»* (TestF)
20 séquences sonores ont été extraites d'une entrevue informelle d'une femme de 30 ans de classe sociale moyenne et vivant en Provence (France).

B- GROUPES DE SUJETS

• *Test «Québec»*:
Les sujets ayant passé le TestQ sont:
· 56 étudiants de l'Université Laval (sujets québécois: SQ);
· 30 étudiants de l'Université de Provence (sujets français: SF);
· 7 phonéticiens de l'Université de Provence (experts: Ex)

• *Test «France»*:
Les sujets qui ont été soumis au TestF sont:
· 26 étudiants de l'Université Laval (SQ);
· 30 étudiants de l'Université de Provence (SF);
· 7 phonéticiens de l'Université de Provence (Ex).

Il est à noter qu'à l'exclusion du groupe des experts qui a été soumis au TestQ et au TestF, les autres groupes ont entendu soit le corpus du TestQ, soit celui du TestF.

C- PROCÉDURE

Les sujets disposaient d'une transcription orthographique sans ponctuation pour chaque séquence du test, les syllabes étant séparées par une barre oblique; - les sujets écoutaient chaque séquence 5 fois (sauf dans le cas des 56 SQ du TestQ qui ont écouté chaque séquence 7 fois). Les répétitions d'une même séquence étaient séparées par un intervalle de 2 secondes, alors qu'une pause de 4 secondes séparait les différentes séquences;
deux tâches étaient demandées aux sujets, à savoir:
· premièrement de souligner les syllabes qui leur semblaient marquées ou mises en relief par rapport à d'autres;
· deuxièmement d'encercler la syllabe la plus proéminente de la séquence.

RÉSULTATS

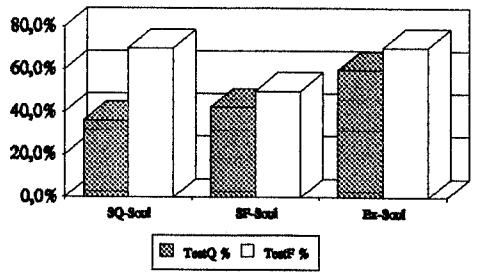


Figure 1 Pourcentages de syllabes soulignées par 15% et moins de sujets, par groupe de sujets et par test.

Les résultats présentés ci-dessous constituent un premier examen des données et ne permettront de vérifer qu'en partie seulement les hypothèses formulées précédemment. À l'aide du test T, nous avons procédé à une série

de comparaisons des moyennes inter-groupes et inter-tests afin de vérifier l'existence de différences significatives entre les trois groupes de sujets quant aux taux de soulignement et d'encerclement, selon la variété linguistique en cause.

La *Figure 1* donne pour chacun des 3 groupes de sujets (Québécois, Français et Experts) et pour chacun des deux tests (TestQ et TestF) le pourcentage de syllabes qui n'ont à peu près pas été perçues comme étant accentuées, soit les syllabes soulignées par moins de 15% des sujets. Cette figure montre des différences statistiquement significatives entre les taux de non-perception ou de non-soulignement dans les deux tests, davantage de syllabes n'étant pas perçues comme accentuées dans la variété française comparée à la variété québécoise. On note également que dans le cas du TestQ, les experts soulignent significativement moins de syllabes que les deux autres groupes, alors que pour le TestF, ce sont les experts et les sujets québécois qui soulignent significativement moins de syllabes que les sujets français. On observe donc des différences entre les variétés linguistiques et les groupes quant au taux de syllabes soulignées. Par contre, étant donné la nature de la tâche assignée dans le cas de l'encerclement où une seule syllabe devait être identifiée, aucune différence significative n'a été observée.

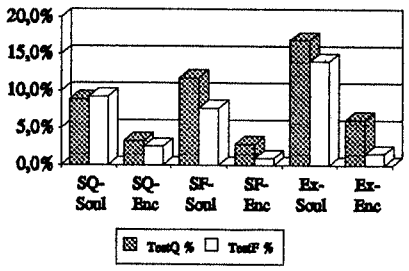


Figure 2 Pourcentages de syllabes soulignées et encerclées par 65% ou plus de sujets, par groupe de sujets et par test.

La *Figure 2* donne la fréquence relative des syllabes soulignées et encerclées par 65% et plus des sujets pour chacun des trois groupes et pour les deux tests. Celle-ci montre bien qu'en général le nombre de syllabes pour lesquelles il y a un accord assez marqué est plus important pour les stimuli produits dans la variété québécoise que pour ceux produits dans la variété de l'Hexagone, bien que seulement 2 des 6 différences soient statistiquement significatives. De plus, les résultats indiquent que les comportements perceptuels des sujets québécois et français sont plus près l'un de l'autre qu'ils ne le sont de celui des experts; ainsi, sur 4 cas possibles, il n'y a aucune différence significative entre les SQ et les SF, alors qu'on observe 3 différences significatives sur 4 possibles entre les experts et les non-experts. On observe encore ici une tendance similaire à celle de la figure précédente en ce qui concerne la différence entre les deux variétés linguistiques et le comportement différencié du groupe d'experts.

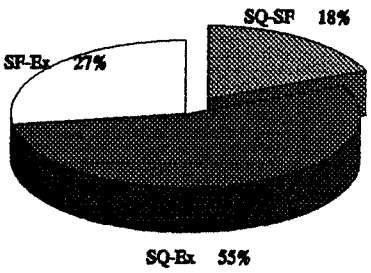


Figure 3 Proportion (en %) des syllabes présentant une différence de 30% ou plus entre les 3 groupes (tâche de soulignement): TestQ.

La *Figure 3* et la *Figure 4* montrent, pour le TestQ et le TestF respectivement, les pourcentages de syllabes qui présentent un écart de 30% et plus entre deux groupes de sujets en ce qui concerne la tâche de soulignement.

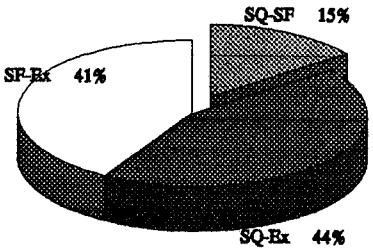


Figure 4 Proportion (en %) des syllabes présentant une différence de 30% ou plus entre les 3 groupes (tâche de soulignement): TestF.

Il ressort clairement de ces deux figures que le taux de syllabes pour lesquelles on observe un écart supérieur à

30% est beaucoup plus faible lorsqu'on compare les sujets non experts entre eux que lorsque ceux-ci sont comparés aux experts. De plus, l'écart entre les sujets français et les experts est beaucoup plus grand lorsque leur propre variété linguistique est en cause (TestF). Ici encore, la variété linguistique ainsi que le degré d'expertise en phonétique semblent jouer un rôle dans les écarts observés.

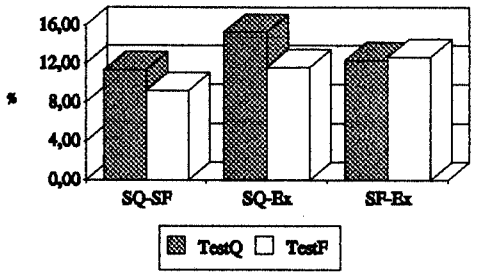


Figure 5 Moyenne des différences entre groupes en pourcentages absolus pour les deux tests: tâche de soulignement.

Lorsqu'on compare les trois groupes de sujets entre eux par une mesure des différences moyennes entre leurs pourcentages respectifs de soulignement et d'encerclement lors des deux tests, on observe les tendances suivantes à partir des *Figures 5* et *6*: sauf en ce qui concerne la comparaison entre les sujets français et les

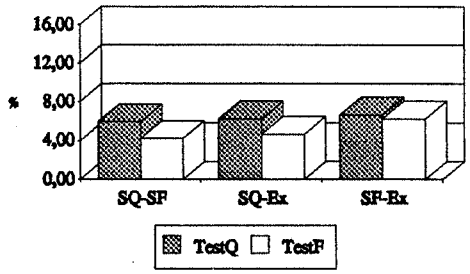


Figure 6 Moyenne des différences entre groupes en pourcentages absolus pour les deux tests: tâche d'encerclement.

experts, les différences moyennes entre les groupes sont significativement plus élevées dans le TestQ que dans le TestF, de sorte que les différences entre les sujets québécois et les sujets français ou entre ces premiers et les experts sont plus élevées lorsque la variété québécoise est en cause, ce quelle que soit la tâche impliquée; de même, pour les deux tâches et les deux tests, les différences entre les sujets non experts (SQ et SF) sont plus faibles que celles que l'on retrouve entre ceux-ci et les experts; enfin, il convient de souligner que, dans le cas du TestF, les différences moyennes sont plus élevées entre les sujets français et les experts qu'entre les sujets québécois et les experts.

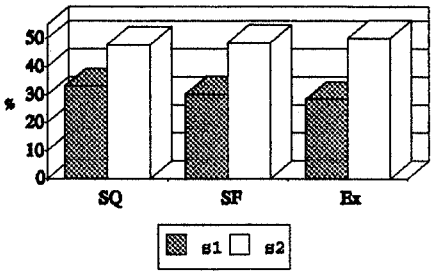


Figure 7 Pourcentages moyens de soulignement pour les premières et deuxièmes syllabes des dissyllabes: TestQ.

Au-delà des différences générales analysées précédemment, nous avons vérifié si les sujets avaient tendance à donner le même poids aux deux syllabes d'unités bisyllabiques, et ce dans les deux variétés linguistiques, en comparant les pourcentages moyens de soulignement. La *Figure 7* et la *Figure 8* montrent que si les deuxièmes syllabes des dissyllabes sont soulignées significativement plus souvent que les premières syllabes dans les deux variétés linguistiques, les premières syllabes sont par ailleurs significativement plus souvent soulignées en français québécois qu'en français de l'Hexagone. La différence de proéminence perçue entre les deux syllabes d'unités dissyllabiques est donc sensiblement plus grande en français de l'Hexagone qu'en français québécois, résultat qui pourrait avoir une incidence sur une perception différenciée du rythme dans les deux variétés linguistiques.

DISCUSSION

Les résultats précédents démontrent que les sujets, fussent-ils experts ou non-experts, sont en mesure de

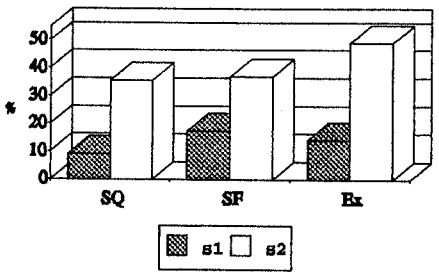


Figure 8 Pourcentages moyens de soulignement pour les premières et deuxièmes syllabes des dissyllabes: TestF.

percevoir divers niveaux de proéminence, comme en témoignent les pourcentages variables de soulignement. De même, aucun comportement opposé n'ayant été observé entre les sujets non experts et les sujets experts, il ressort que les données obtenues auprès de l'un ou l'autre de ces groupes sont tout aussi fiables l'une que l'autre. En ce qui concerne la première hypothèse, il semble bien que la proéminence, réalisation phonétique de l'accent, est bel et bien perçue comme un continuum, si on se fonde d'une part sur le fait qu'un très faible nombre de syllabes est associé à des jugements catégoriques (0% ou 100%) et, d'autre part, sur la distribution des jugements obtenus dans le cas des deux syllabes des unités dissyllabiques. De plus, tel que formulé dans la troisième hypothèse, les résultats démontrent que, bien qu'il existe un certain nombre de différences entre les deux groupes de sujets non experts, les différences les plus importantes ont été obtenues en comparant les experts aux non-experts, ce qui laisse supposer l'existence d'au moins deux types différents de stratégies perceptuelles. Cependant, contrairement à l'hypothèse voulant que l'appartenance à une communauté géo-socio-linguistique allait avoir un impact sur la perception des degrés de proéminence, il semble plutôt que c'est la variété linguistique elle-même qui exerce une telle influence.

Bien qu'il nous faille approfondir ces résultats par une série d'études complémentaires relatives aux facteurs acoustiques, aux stratégies individuelles, à la caractérisation linguistique et pragmatique des séquences, etc., ceux-ci sont tout de même déjà révélateurs de l'importance qu'il y a de concevoir l'analyse de l'accent dans une perspective dynamique, plutôt que d'après un ordre du tout ou rien, et dans une optique théorique qui explique la variation inter-linguistique et inter-groupe si on veut rendre compte de ses véritables nature et rôle.

RÉFÉRENCES

Abramson A.S. et L. Lisker L. (1970). «Discriminability along the voicing continuum: cross-language tests», *Proceedings of the 6th International Congress of Phonetic Sciences*, Prague, Academia, pp. 569-573.

Allen, G.D. (1972). «The location of rhythmic stress beats in English: an experimental study», *Language and Speech*, vol. 15:1, pp.72-100.

Bolinger, D. L. (1958). «On intensity as a qualitative improvement of pitch accent», in *Lingua*, vol. 7, pp.175-182.

Boudreault, M. (1968). *Rythme et mélodie de la phrase parlée en France et au Québec*, Québec, Presses de l'Université Laval.

Bruneau, Ch. (1931). *Manuel de phonétique pratique*, 2e édition, Paris, Berger-Levrault.

Delattre, P. (1938). «L'accent final en français: d'intensité, de hauteur, de durée», in *French Review*, vol. 12, no. 2, pp. 141-145.

Di Cristo, A. (1978). *De la microprosodie à l'intono-syntaxe*, Thèse de doctorat, Université de Provence, Aix-Marseille.

Di Cristo, A. (sous presse). «Intonation in French», in D. Hirst et A. Di Cristo [dir.], *Intonation systems. A survey of twenty languages*, Cambridge, Cambridge University Press.

Di Cristo, A. (1981). «Aspects phonétiques et phonologiques des éléments prosodiques», *Modèles linguistiques*, vol. 3:2, Lille, Presses universitaires de Lille, pp. 24-83.

Dittman, A.T. et L.C. Wynne (1961), «Linguistic techniques and the analysis of emotionality in interviews», *Journal of abnormal and social psychology*, vol. 63, pp. 201-204.

Fonagy, I. (1979). «L'accent en français: accent probabilitaire», in I. Fonagy & P.R. Léon [dir.], *L'accent en français contemporain*, Studia Phonetica, vol. 15, Montréal, Didier, pp. 123-233.

Fry, D.B. (1958). «Experiments in the Perception of Stress», in *Language and Speech*, vol. 1, pp. 126-152.

Guaïtella, I. (1991). *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, thèse de doctorat, Université d'Aix-Marseille I.

Hirst, D. et A. Di Cristo (1984). «French Intonation: a parametric approach», in *Die Neueren Sprachen*, vol. 83, no.5, pp. 554-569.

Lieberman, P. (1960). «Some acoustic correlates of word stress in American English», in *Journal of the Acoustical Society of America*, vol. 32, pp. 451-454.

Lieberman, P. (1965). «On the acoustic basic of the perception of intonation by linguists», *Word*, vol. 21, pp. 40-54.

McDowall, J.J. (1974). «The reliability of ratings by linguistically untrained subjects in response to stress in speech», *Journal of Psycholinguistic Research*, vol. 3, pp. 247-259.

Mertens, P. (1991). «Local prominence of acoustic and psychoacoustic functions and perceived stress in French», *Actes du XIIe Congrès international des sciences phonétiques*, vol. 3, pp. 218-221.

Morton, J. et W. Jassem (1965). «Acoustic correlates of stress», in *Language and Speech*, vol. 8, pp. 159-181.

Paradis, C. (1993). «L'accent: réflexions et notes de lecture». *Dialangue*, , vol. 4, avril 1993, pp.9-19.

Paradis, C. et D. Deshaies (1990). «Rules of stress assignment in Québec French: Evidence from perceptual data», *Language Variation and Change*, vol.2, pp. 135-154.

Paradis, C., Dolbec, J., et M. Brousseau (1993). «Enjeux sociolinguistiques en synthèse de la parole: variétés linguistiques et intellligibilité du message», *Revue québécoise de linguistique*, vol. 22:2, pp.13-36.

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 Phonetic Sciences*, La Haye, Mouton, 1961, pp. 735-748.

Rossi, M., Di Cristo, A., Hirst, D., Martin, Ph. et Y. Nishinuma (1981). *L'intonation: de l'acoustique à la sémantique*, Paris, Klincksieck.

Singh, S. et J.W Black (1966). «Study of twenty-six intervocalic consonants as spoken and recognized by four groups», *J.A.S.A.*, vol. 65 (supp.1), ss11.

Vaissière, J. (1980). «La structuration acoustique de la phrase française», in *Ann. Scu. Norm. Sup. Pisa*, vol. III, no. 10, pp. 529-560.

Warren, R. et L. Santerre (1979). «Les paramètres acoustiques de l'accent en français montréalais», in I. Fonagy & P.R. Léon [dir.], *L'accent en français contemporain*, Studia Phonetica, vol. 15, Montréal, Didier, pp.53-63.

CARACTÉRISATION DES VARIATIONS INDIVIDUELLES DU CONTOUR DE FREQUENCE DU FONDAMENTAL OBSERVÉES DANS DES PHRASES LUES EN ANGLAIS.

JACQUELINE VAISSIÈRE
Institut de Phonétique, CNRS-UA 1027
75005 Paris, France
et
ATR Interpreting Telephony Research Laboratoires
Kyoto, Japon.

Résumé:

Le but de cette présentation est de décrire une méthode de caractérisation des courbes du fondamental (Fo) à trois étapes, permettant de capturer les différences individuelles. Premièrement, la phrase est segmentée sur la base du passage de Fo par la ligne de base (BL) en "unités de sens", nommés HP (par analogie avec "hat-pattern"). Une "unité de sens" est composée des mouvements suivants: [montée + (plateau P) + descente + (Montée de continuation, Rc)]. Deuxièmement, la distance acoustique entre HPs successifs est estimée selon les critères suivants: (i) présence ou absence d'une pause, précédée ou non, sur la dernière syllabe avant la pause, par Rc ou une valeur extra-basse, (ii) interruption de la tendance à décliner régulièrement: soit un rehaussement ("resetting") ou un nivellement, (iii) un pic de Fo plus élevé après la frontière et/ou un abaissement du pic de Fo avant la frontière. Troisièmement, le degré de jointure entre les mots à l'intérieur d'un HP est estimé selon l'importance de la rupture de la continuité du Fo sur P: présence d'un Rc, ou pic de Fo sur l'une des syllabes accentuées. La division de la phrase en HPs est sensiblement la même pour les locuteurs. Les différences de degrés de frontières marqués entre HPs successifs, et les sous-divisions des HPs, et surtout les procédés les plus fréquemment employés permettent de comparer les productions des locuteurs. L'étude du Fo de 1000 phrases nous a permis de redécouvrir la validité du principe du HP pour la parole lue, et de mettre en évidence des "habitudes prosodiques" (séquences particulières de mouvements de Fo) propres à chacun des locuteurs.

Mots-clés: *prosodie, anglais, différences entre locuteurs, modélisation*

INTRODUCTION

Les variations de fréquence du fondamental (Fo), durée et intensité, sont contrôlées en partie par des règles d'origine linguistique et qui sont différentes dans les langues à tons, les langues à accent de mot contrastif, les langues où la position de l'accent lexical est fixe par rapport aux frontières de mot, ou dépend de la structure morphémique, ou de la catégorie grammaticales des mots, etc. De plus, les marques prosodiques d'intonation dépendent de la coexistence éventuelle de marqueurs morphologiques ou de transformations syntaxiques (pronom interrogatif ou de l'inversion du sujet pour marquer une question, par exemple).

Les variations dépendent également de phénomènes plus généraux, relevant des particularités du système de perception ou de production (Collier, 1984; Vaissière, soumis par publication), indépendants de la langue en question. Le cantonnement à des explications relevant de la linguistique ou à la phonologie particulière d'une langue peut aboutir à un obscurcissement des problèmes (Voir Garde, p. 5).

Plus on dispose de données sur diverses langues, et plus on est frappé par les similarités prosodiques entre les langues (Vaissière, 1982): on peut donner pour exemples l'utilisation des pauses pour la segmentation, l'adaptation faite à d'autres langues de la célèbre formule de Dennis Klatt pour la génération de la durée en anglais, et la tendance de Fo à décliner tout au long des phrases, dans les langues à tons (Meyers, 1974) ou autres (Cohen et t'Hart, 1967), Fo plus élevée après les consonnes sourdes, etc. Selon notre hypothèse, des propriétés acoustiques, sans doute incontrôlées au départ, d'origine physiologique, sont devenues des réalités psychologiques: la tendance à la déclinaison devient une caractéristique nécessaire pour une synthèse correcte des phrases déclaratives (Abramson, 1978 sur le Thaï), et elle est prise en compte par l'auditeur pour la comparaison de la hauteur des pics (Pierrehumbert, 1979, sur l'anglais). Ces tendances, dont on peut faire l'hypothèse qu'elles sont les mêmes pour toutes les langues, peuvent être phonologisées ou non dans une langue donnée..

Malgré l'abondance des recherches, il manque toujours un modèle de notation des courbes de Fo permettant de mettre en valeur les différences entre les langues. Le problème est qu'il n'existe pas de modèle unique pour une langue donnée, et plus une langue est étudiée, et plus les modèles abondent et se contredisent, du moins, en partie. Notre approche de description est très fortement marquée par nos expériences sur l'utilisation de la prosodie en reconnaissance, où on se trouve confronté par les différences entre locuteurs. Nous ne croyons pas au "modèle miracle", satisfaisant à la fois les besoins de la synthèse (codage des effets combinés du contexte, de la syntaxe et du rythme), ceux de la reconnaissance (informations à extraire automatiquement, Vaissière, 1985), des linguistes (données permettant de renforcer ou d'infirmer telle ou telle hypothèse), de l'orthophoniste et du professeur de langue (notation des erreurs prosodiques), etc. D'autres

méthodes sont actuellement développées: le système TOBI pour l'anglais, basé à la fois sur l'inspection visuelle de la courbe de Fo et l'audition des phrases, actuellement en cours de développement et qui représente ce sur quoi un groupe d'experts reconnus en prosodie a réussi à s'entendre, le système automatique de stylisation des courbes de Fo développé à Aix, (Hirst et Di Cristo (1993), etc. Il reste à tester leur capacité à décrire les différences entre les locuteurs.

Le français, avec ses divers degrés de continuation et de finalité marqués par Fo, et d'allongements finals (Delattre, 1962) illustre comment une langue peut jouer des paramètres (Fo, durée) pour marquer différents degrés de frontière. En anglais, les paramètres prosodiques sont fortement influencés par la réalisation des accents de mots, mais de nombreux articles ont montré que la différence de degré entre accents de mots, la réalisation d'allophones initiaux, et l'allongement final, contribuaient aussi à marquer des frontières de degrés différente entre les mots. L'aspect démarcatif de Fo est-il plus difficile à mettre en évidence en anglais qu'en français? La structure prosodique est-il le même en anglais qu'un français?

1. CORPUS, LOCUTEURS ET MATÉRIEL

Nous avons utilisé une partie du corpus sur l'anglais disponible sur le réseau informatique de la compagnie ATR, à Kyoto, 200 phrases prononcée par 5 locuteurs (4 britanniques et 1 américain). L'affichage du spectrogramme, de la segmentation, du Fo et l'écoute d'une liste de phrases préalablement définie est possible en un temps très rapide (utilisation du logiciel Xwaves), et ces facilités ont rendu possible l'étude d'un grand nombre de phrases, dans d'excellentes conditions.

2. DESCRIPTION

Rappel de la notion de hat-pattern

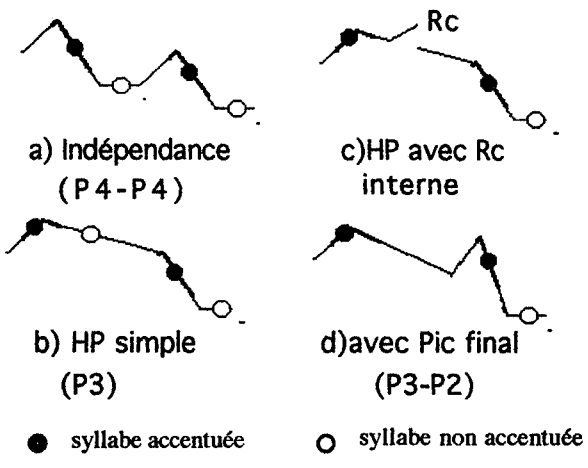


Figure 1a et b: Principe du groupement par hat-pattern: la séquence [montée+descente + montée+descente] correspondant à deux mots (en a) est réduite en la séquence [montée + descente] (en b)., correspondant aux 2 mots regroupés.

Figure 1c et 1d: Décomposition d'un hat pattern: par une montée de continuation (c) ou un pic sur le plateau (d). La ressemblance avec les patterns français a été indiquée.

Notre description est basée sur le principe bien connu du HP (expression reprise de l'école hollandaise de prosodie, Cohen et t'Hart, et appliquée à l'anglais par Maeda, Maeda, 1975). Le principe est fort simple. La figure 1a schématise la réalisation de deux accents successifs (schématisation inspirée de Delattre, 1962). La figure 1b illustre le principe de regroupement par HP: le premier mot tend à recevoir un mouvement montant (ton haut ou parfois montant sur la syllabe accentuée), et le second une descente de Fo. Les syllabes inaccentuées centrales tendent à rester sur le registre Haut, et la forme générale de la courbe de Fo est celle d'un chapeau (procédé 1). Une autre réalisation consiste à réaliser les deux mouvements opposés dans le premier mot, et à désaccentuer le second (procédé 2).. Le nombre de HPs dans une phrase correspond grosso modo à sa division en groupes de sens. Ces deux principes de regroupement nous semblent relever de principes éthologiques (voir Vaissière, en préparation) et seraient en conséquent des tendances universelles (à démontrer).

La figure 4 illustre des différences entre locuteurs. le loc 1 (L1) a regroupé "the candidate felt" et " trapezoidal badge", L1) par le procédé 1, et les deux autres locuteurs ont utilisé le procédé 2. Quand l'accent lexical du premier mot est situé après les deux premières syllabes du mot ("trapezoïdal"), la montée tend à correspondre à la première syllabe du mot (pour les variations entre locuteurs sur ce point, voir Vaissière, 1985- la même tendance a été observée en Espagnol, Vaissière 1983). (Seuls 3 locuteurs sont illustrés sur la figure 4 par manque de place).

2.A. Division "manuelle" en HP.

La figure 4 illustre le résultat d'un décomposition visuelle en HP par passage par BL (même phrase prononcée par trois locuteurs, L1, L2 et L3, ([] indique la retour à la BL).

Par exemple, pour L1, les regroupements pourraient se décrire de la façon suivante:

"The candidate felt [BL+Rc] that a trapezoidal badge [BL+Rc] would have more impact [BL+Rc] than [BL] the usual [BL] rosette" (6 HPs))

2.B. Division automatique en HP

L'automaticité est un fait hautement désirable pour l'étiquetage de grandes bases de données. Comme la pente générale de BL peut changer d'une phrase à l'autre pour un même locuteur (en fonction de sa durée, de son rang dans la liste, de la fatigue du locuteur, etc.) et même à l'intérieur de la phrase, on a cherché un critère de détection automatique des passages par BL pouvant s'appliquer sur la base d'une seule phrase. Cette recherche s'est avérée vaine.

La solution qui nous a paru la meilleure, mais qui nécessite quand même une vérification manuelle des

résultats, est la suivante et s'opère en deux temps. Premièrement, à partir du Fo de la première syllabe, on essaie de tracer une droite, la plus longue possible, qui suive les mouvements de Fo de la suite sans les croiser (figure 2a). Un croisement indique que la BL est atteinte (par exemple, l'intersection de la courbe de Fo avec les droites 3, 5, 6 et 7 sur notre schématisation) , et on recommence à partir de ce dernier point, qui devient le premier point du tronçon du HP suivant. Le nombre de passage par BL permet de compter les HPs. Dans un deuxième temps, en partant de chaque passage par BL, on trace une droite vers la gauche pour atteindre les valeurs plus basses du début du HP, sans jamais croiser la courbe de Fo (prise en compte du rehaussement). Les tronçons de lignes pleines (étape 1) et les tronçons de lignes pointillées (étape 2), à la figure 4, illustrent schématiquement ce qu'aurait pu donner la segmentation automatique avec un tel algorithme.

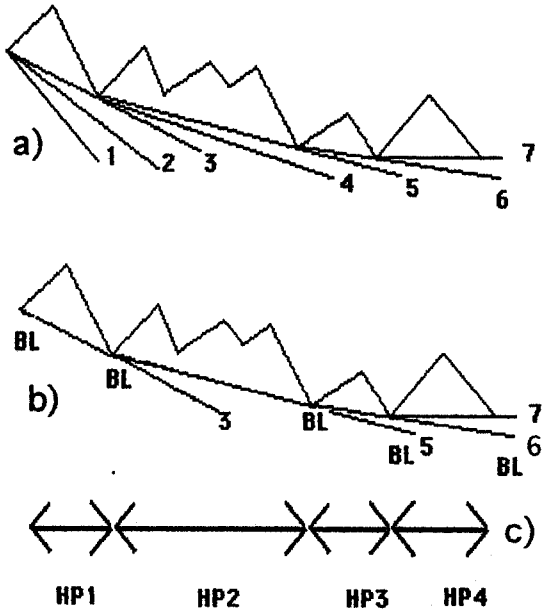


Figure 2: Première étape de la détermination automatique des passages par la ligne de base BL

La position exacte de la frontière entre deux HP tient compte de la limite des phonèmes. Si le passage par BL se fait entre deux voyelles, ou à la fin d'une voyelle ("felt" , "badge", "impact", "visual", "rosette", pour L2, Fig. 4), on coupe à l'endroit exact du passage par BL; sinon, on coupe à la fin de la voyelle ("felt", "badges", "impact" pour L1, "candidate", "badge", "impact" pour L3), et ce "retard" est dû à la présence d'un Rc (règle identique en français).

2.C. Marquage de degrés de frontière entre HPs successifs

La phrase anglaise peut recevoir soit une structure que nous interprétons comme parallèle, où les coupures entre les HPs semblent être d'égale importance (figure 3a), soit une structure marquée (figures 3b à 3f), comme dans notre description traditionnelle pour le français. Le degré de frontière entre les HPs peut être varié par le locuteur à l'aide d'un certain nombre de procédés classiques, que les divers articles sur la prosodie ne cessent d'illustrer (il n'est pas possible d'en faire une liste ici). Ces procédés incluent l'insertion d'une pause (fig. 3c; Grosjean et al., 1979), (généralement accompagnée d'un allongement prépausal de la syllabe finale), la "cassure" de la continuité de BL par un nivellement (fig. 3b), ou un rehaussement (en anglais, "Resetting", Maeda, 1975, Cooper et Sorensen, 1981; en japonais, c'est une des composantes principales du modèle de Fujisaki, Fujisaki & al, Sudo (1971) souvent accompagné d'un rehaussement du plateau (fig. 3d), ou encore le pic sur un HP peut être abaissé par montrer sa dépendance avec le HP précédent (Fig. 3e) ou élevé pour au contraire, montrer son indépendance (Fig. 3f). Le fondamental sur la dernière syllabe avant la frontière peut recevoir une montée de continuation (un crochet ou "hook" en anglais selon la description de Delattre, 1962) ou encore (plus rarement) une valeur de Fo plus basse (voir le choix opposé des locuteurs 2 et 3 pour le mot "candidate", Fig. 4). Tous ses procédés sont relativement indépendants: on peut avoir un "resetting" ou une montée de continuation sans l'insertion de pause, par exemple et des études de perception sont nécessaires pour comparer leur efficacité perceptive à segmenter et hiérarchiser les groupes de sens dans la phrase.

2.D Subdivision d'un hat-pattern et différences entre locuteurs.

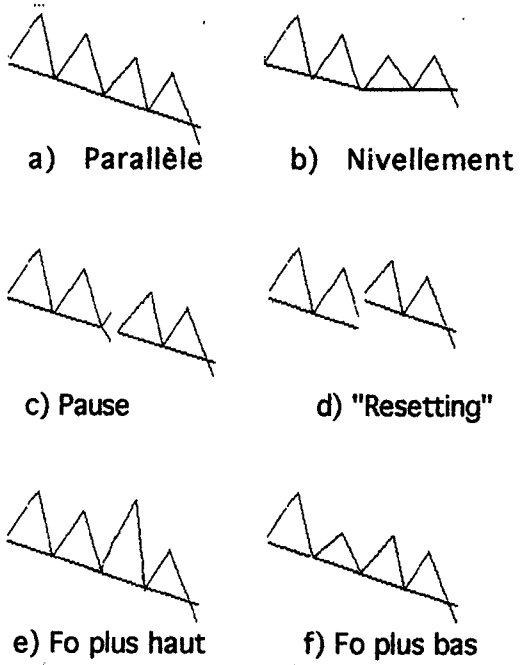


Figure 3: Quelques procédés classiques utilisés pour l'augmentation du degré de séparation entre les deuxième et troisième HPs.

Alors que la division en HPs a été sensiblement la même pour tous les locuteurs (voir Lea, 1974), les variations de Fo à l'intérieur d'un même HP ont montré de grandes différences individuelles. Par exemple, dans la suite (adjectif + nom), "trapezoidal badge", on observe une montée de Fo durant la consonne initiale sourde de "trapézoïdal" pour L1, L2 et L3. La descente est réalisée durant la syllabe accentuée du second mot "bAdge" pour L1, mais après la syllabe accentuée du premier mot pour L2 et L3.

Nous savons par expérience que ces différences ne sont pas aléatoires, elles sont la conséquence d'une ou plusieurs tendances sous-jacentes en conflit, que chaque locuteur tend à régler de façon individuelle. L'analyse visuelle des 1000 phrases montre de nombreuses régularités dans les différences entre locuteurs. On citera deux exemples:

a) la montée de Fo peut occuper toute la plage, et l'amplitude de la descente est alors inférieure à celle de la montée (Loc. L1) ou sensiblement égale (L2). Ou alors, la montée initiale n'occupe qu'une part réduite de la plage, et on observe un pic sur le plateau et une descente d'amplitude plus importante (L3).

(b) Le premier et le dernier HPs dans une phrase peuvent être plus ou moins influencés par leur position dans la phrase (pattern éventuellement "plus montant" en début de phrase pour L1).

3. DISCUSSIONS ET CONCLUSION

Bien que la ressemblance puisse être l'effet du hasard, les schémas typiques des trois locuteurs anglais rappellent ceux des locuteurs français (Vaissière, 1975), P3-P4 , P3-P2 , et P2 pour L1, et L2 et L3, respectivement.

Ces répétitions de patterns identiques d'une phrase à l'autre chez un même locuteur contribuent très certainement à l'impression du style particulier à chaque locuteur. Le degré différent des perturbations de ces schémas en début et en fin de phrase est peut-être dû à un poids différent à accorder au constituant HP dans sa combinaison avec les constituants de rang supérieur, GS et inférieur (Vaissière, 1992).

Le passage de Fo par la BL est donc une base essentielle dans notre description. Ce passage nous semble plus important que les détails de la courbe de Fo. Le retour à la BL semble avoir une signification particulière du point de vue physiologique, une sorte de remise à zéro, simulant la fin d'un effort physiologique. Les données sur la perception des courbes de Fo semblent aller dans cette direction: une chute de Fo ne donnera l'expression de finalité que si le Fo atteint la BL (Collier, 1984), mais ces résultats demandent à être confirmés sur des locuteurs de plusieurs langues.

Cette description n'est qu'une description de plus à ajouter à la liste déjà longue des descriptions offertes pour l'anglais. Elle a pour avantage d'être inspirée essentiellement par les données elles-mêmes. Une première inspection de la base de quelques données disponibles sur le Japonais à ATR (avec l'aide appréciée de nos collègues et d'Hélène Valbret) et à travers la littérature (Fujisaki et Sudo, 1971; Kubozono, 1993) nous laisse à penser que les principes décrits ci-dessus pour le regroupement des HP sont également valables pour le japonais (cf. les exemples illustrés dans le livre de Kubozono). Nous pensons que des progrès en prosodie passe désormais par une accumulation de données sur de très nombreuses langues (voir, par exemple, Touati, 1987, pour une comparaison suédois-français).

Cette étude est très incomplète dans le sens où elle concerne uniquement les pauses et les variations de Fo, sans considération des variations d'intensité et surtout de durée. La structure temporelle à elle seule (par un allongement de voyelle finale) peut être responsable de la perception de ruptures (Duez, 1994 pour le français). Des études statistiques sur les structures temporelles des phrases sont une tradition à ATR (Sagisaka, 1992; Campbell, 1991). Il serait intéressant de comparer nos résultats avec les "courbes de durée", calculées par Nick Campbell sur les mêmes 1000 phrases. L'étude de la corrélation entre durée et intensité, durée et Fo s'est révélée négative en Japonais (Sagisaka, 1992), alors que l'amplitude et la durée semblent très corrélées en anglais (Fry, 1955). Il faudra peut-être, dans une étude des rapports entre durée, et Fo en anglais pour le marquage des frontières, tenir compte du fait que ce ne sont pas tant les valeurs de Fo sur les voyelles accentuées ou les tons qui importent, mais le passage de Fo par BL .

REMERCIEMENTS

Je tiens beaucoup à remercier la compagnie ATR qui m'a offert un havre de recherche durant les étés de ces deux dernières années, en particulier Messieurs Sagisaka et Campbell, dont j'ai beaucoup appris.

RÉFÉRENCES

Abramson, A.S., (1978), "Static and dynamic acoustic cues in distinctive tone", Language and Speech, 21, 319-325.

Campbell, N, (in press), "Automatic detection of prosodic boundaries in speech" accepté for publication in Speech Communication.

Campbell, N., (1991), "Segment duration in a syllabic frame", J. Phon. 19, 37-47.

Cohen, A., and t'Hart, (1967), "On the anatomy of intonation", Lingua 19, 177-192.

Collier, R., (1984), "Some physiological and perceptual constraints on tonal systems", dans EXPLANATIONS FOR LANGUAGE UNIVERSALS, Butterworth & al (éditeurs), Mouton, New York, Berlin, Amsterdam.

Cooper, W.F., & Sorensen, J.M., (1981), "FUNDAMENTAL FREQUENCY IN SENTENCE PRODUCTION, New-York: Springer.

Delattre, P., (1962), "Comparing the prosodic features in English, German, Spanish, and French", Int. Rev. Applied. Linguistics., I, 193-210.

Duez, D, (1994)., "Acoustic correlates of subjective pauses", J. Psychological Research, Vol. 22, N°. 1, 21-38.

Fry, D.B., (1955), "Duration and intensity as physical correlates of linguistic stress", JASA, 27, 765-768.

Fujisaki, H. and Sudo, H., (1971), "A generative model for the prosody of connected speech in Japanese", Ann., Res. Eng. Inst. Logopedics Phoniatrics, Univ. Tokyo 30, 75-80.

Grosjean, F., Grosjean, L., and Lane, H., (1979), "The patterns of silence: performance structures in sentence production." Cognitive Psychology, 11, 58-81.

Hirst, D; et Di Cristo, A., (à paraître), INTONATION SYSTEMS: A SURVEY OF TWENTY LANGUAGES, Cambridge University Press. Hirst, D; et Di Cristo, A., (à paraître)

Kubozono, H., (1993), THE ORGANISATION OF JAPANESE PROSODY, Studies in Japanese Linguistics, Series Edition, Kurisio Publishers.

Lea, W.A., (1974), " Prosodic aids to speech recognition: a summary of results to date", Sperry Univac Technical Report, No. PX 11087.

Maeda, S., (1976), A CHARACTERISATION OF AMERICAN ENGLISH INTONATION, Ph. D. Thesis MIT.

Meyers, L., (1974), "Tone pattern in Hausa: a re-analysis of Hausa downdrift, UCLA Working papers in Phonetics, 27, 47-62.

Pierrehumbert, J., (1979), "The perception of fundamental frequency declination", JASA, 66, 363-369.

Sagisaka, Y., (1992), "On the modelling of segmental duration control", dans SPEECH PERCEPTION, PRODUCTION AND LINGUISTIC STRUCTURE, édité par Y. Tokhura, E. Vatikiotis-Bateson, et Y. Sagisaka, Ohmsta IOS Press, Amsterdam, Washington, Oxford.

Touati, P., (1987), STRUCTURES PROSODIQUES DU SUEDOIS ET DU FRANCAIS, Travaux de l'Institut de Linguistique de Lund, Lund University Press.

Vaissière, J., (1982), "Language-independent prosodic features", in PROSODY, MODELS AND MEASUREMENTS, édité par A. Cutler et R. Ladd, Springer-Verlag, 53-66.

Vaissière, J, (1985), " The use of prosodic parameters in automatic speech recognition', dans COMPUTER, SPEECH AND LANGUAGE, édité par F. Fallside et W, Woods, Prentice Hall International.

Vaissière, J., (1992), "Rhythm, accentuation and final lengthening in French", in Music, Language, Speech and Brain, édité par J. Sundberg, L. Nord, R. Carlson, Wenner-Gren International Symposium Series, Vol. 59, 108-120.

Vaissière, J., "Explanation of prosodic characteristics: ethological, articulatory and acoustico-perceptive constraints, and psychological associations" soumis pour publication dans le livre dédié à Max Waskopf.

Figure 4: Fréquence du fondamental de la phrase:
"The candidate felt that a trapezoidal badge would have more impact that the usual rosette."

Sont indiqués de haut en bas:
1) les patterns stylisés,
-2) le signal
-3) la courbe du fondamental,
4) les lignes pleines quasi-horizontales indiquent les tronçons de la ligne de base , et les lignes pointillés marquent le "resetting" ou rehaussement de la ligne de base après une frontière,

On observera les différences inter-locuteurs suivantes

L1:. Tendance P3-P4. à tendance Rc: légère Rc à la fin des HP ("felt", "badge", "impact"), et à la fin des mots à l'intérieur des HP: "candidate", "trapezoidal", "visual").
Les fluctuations de Fo sont dominées par la montée durant le premier mot de chaque HP.

L2: Tendance P3-P2. Renforcement du premier pic de la phrase. Rarement de Rc chez ce locuteur, mais une valeur extra-basse suivie d'un resetting marque la continuation."
Fort contraste d'intensité entre les syllabes accentuées

L3: Tendance P2. avec Rc. Les fluctuations de Fo sur le plateau sont dominées par les pics .sur une syllabe accentuée.

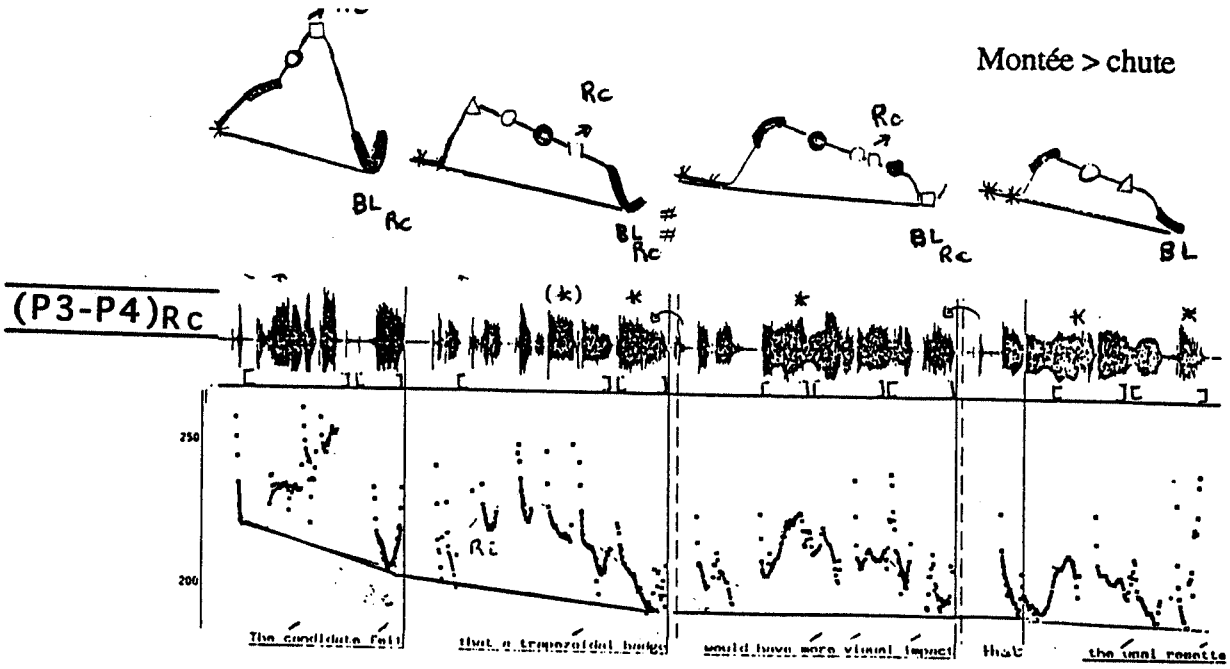
Mot grammatical

Syllabe initiale d'un mot lexical (accentuable mais non accentuée)

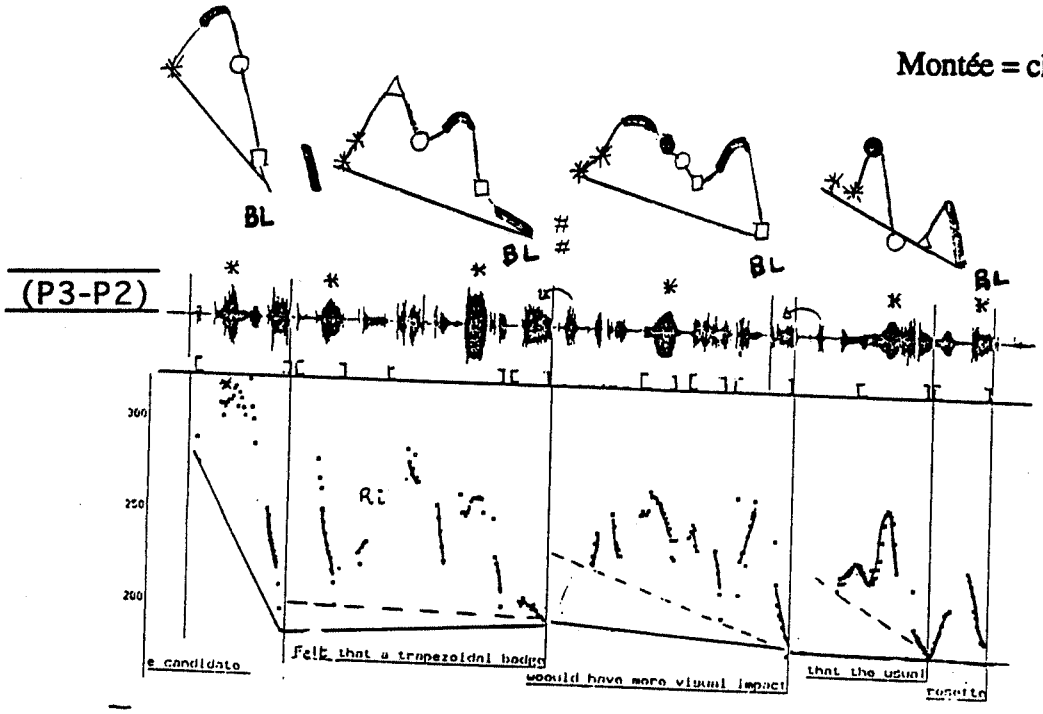
Syllabe finale de mot (non accentuée)

Syllabe accentuée sans mouvement de Fo apparent

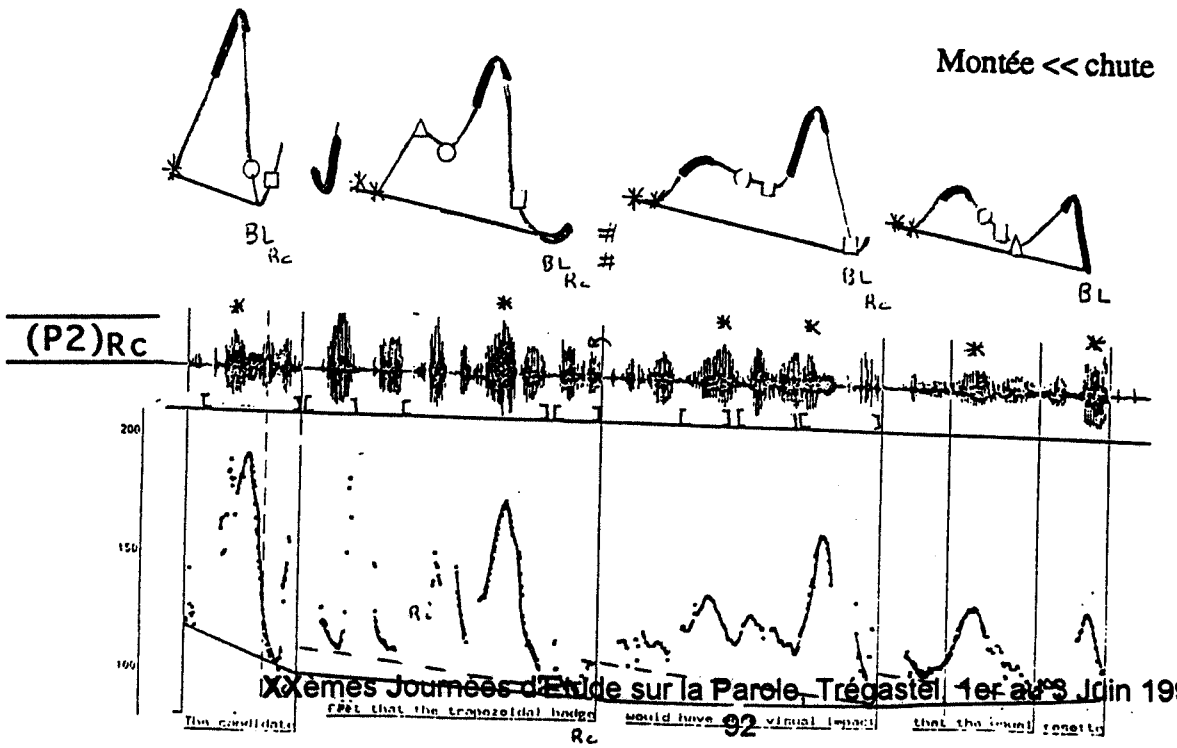
Syllabe accentuée avec mouvement de Fo apparent



Loc 1



Loc 2



Loc 3

SESSION AFFICHEE : A1

PRODUCTION ET SYNTHESE

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

VERS UNE PHONÉMISATION AUTOMATIQUE DES SIGLES

Philippe BOULA de MAREUIL

LIMSI-CNRS BP 133 F-91403 Orsay Cedex
e-mail : mareuil@limsi.fr

Résumé

Les sigles sont aujourd'hui de plus en plus utilisés. Nous proposons dans ce papier des règles pour que, face à un sigle, une synthèse vocale sache s'il faut l'épeler ou le lire, et de quelle façon. Notre analyse repose sur le nombre de lettres qui constituent le sigle, sur la répartition en consonnes et voyelles, mais aussi sur l'opposition entre consonnes continues et momentanées. Elle a donné lieu à une réalisation informatique, où la segmentation des mots isole la ponctuation.

Mots-clés : *sigles, acronymes, siglaison, synthèse vocale*

INTRODUCTION

Les sigles sont de plus en plus employés dans la langue courante contemporaine, pour désigner des organismes administratifs, des associations (même les plus petites), des groupements politiques, des sociétés industrielles ou commerciales, des concepts scientifiques, des pays, des instruments, etc : on ne parle plus de "vélocipède" ni de "vélo tous terrains", mais de "VTT"; "RU" a supplanté "Resto U" et "Restaurant Universitaire". Le sigle acquiert même valeur légale et peut être protégé comme une marque de produit [Doppagne 91]. Quelle que soit la page de journal que l'on se mette à lire, on est immédiatement confronté à cet élément nouveau de notre époque. Les écrivains ne dédaignent plus y recourir, et "ovnis" au pluriel figurait dans l'édition 93 de la dictée de B. Pivot. Les Latins avaient déjà usage de sigles, dans des expressions telles que "SPQR" ("Senatus PopulusQue Romanus"), et le français présente des formes semblables dès le XVIème siècle (avec "RPR" pour "Religion Prétendument Réformée"); pourtant *Le Grand Robert* ne parlait en 1969, au sujet des sigles, que de signes abréviatifs, en faisant allusion aux symboles.

Cas particulier de l'abréviation, la siglaison sert à raccourcir des mots en ne retenant que les lettres ou syllabes initiales de mots composés (ou syntagmes figés). Face à un sigle, 2 attitudes sont donc possibles : le lire comme s'il s'agissait d'un mot, ou l'épeler. Le *ou* n'est d'ailleurs pas exclusif : pour "ONU", par exemple, 2 oralisations sont attestées, même si /ɔny/ tend à s'imposer - en témoigne la formation par suffixation de l'adjectif "onusien".

L'étude de l'oralisation des sigles nous a paru intéressante au moins à double titre : sur le plan théorique - linguistique -, d'abord, on touche là à la base du système phonologique, même si celui des sigles est incomplet (/ŋ/, par exemple, n'y figure pas). Restant dans un domaine où l'on peut quasiment être exhaustif, on teste directement l'importance de la classification des phonèmes. Sur le plan pratique, ensuite, il est difficile de reconnaître "PDU", lorsqu'une synthèse vocale lit /pdy/. Inversement, il est difficile de reconnaître "SOCRATE" épelé, quand on est habitué à l'entendre lu. Or peu de règles ont été mises en application concernant ce sujet : ainsi dans la synthèse vocale VoxBox , il en existe uniquement pour les sigles ne comprenant pas de voyelles (ex. "RG" - voir "Hergé" -), qui sont logiquement épelés.

VoxBox a été mise au point par le KTH de Stockholm [Carlson et al. 82]. Pour développer et tester nos règles, nous l'avons utilisée dans sa version 3.3 pour ordinateur compatible PC, commercialisée par la société Infovox.

1. THÉORIES ACTUELLES

1.1. DÉFINITIONS

Certains parlent parfois d'*acronyme* quand une prononciation syllabique est possible, réservant alors le terme de *sigle* aux groupes de lettres ne constituant pas de mot syllabique et aboutissant à une prononciation alphabétique. Mais il n'y a pas consensus sur ces définitions : d'autres, en effet, utilisent le terme *acronyme* (du grec *acros* "extrême") pour désigner des sigles formés sur la première syllabe des mots qu'ils réduisent (ex. "USINOR") et également des mots constitués par le début d'un mot et la fin ou la totalité d'un autre (ex. "restoroute", sur "restaurant" et "autoroute") [Béchade 92]. Et l'anglais qui ne connaît que le terme "acronym" ne fait pas cette distinction.

Une graphie consistant à faire suivre chaque lettre par un point abréviatif, dans les sigles épelés - ce qui n'est plus nécessaire dans les sigles lus -, permettrait également de lever l'ambiguïté. Mais elle est partout en voie de disparition, surtout si le sigle est long et d'une grande notoriété (ex.

"UNESCO", que l'on peut même trouver écrit "Unesco").

Divers facteurs, notamment socio-linguistiques, sont à l'origine de cette variabilité : la familiarité que l'on a avec le sigle, le mépris que l'on veut montrer envers ce qu'il représente et l'amusement que peuvent provoquer les sonorités (aussi "PMU" est-il parfois lu /pmy/). L'homonymie joue également un rôle : ainsi, le "Service Obligatoire du Travail", que les Allemands avaient instauré pendant l'Occupation, qui donnait "SOT" (oralisé comme "sot" par les Français) a au bout de quelques jours été remplacé par le "Service du Travail Obligatoire" ("STO").

Les études actuelles tentent d'expliquer ces mécanismes; elles sont très intéressantes sur le plan linguistique, mais notre propos ici est de trouver des règles permettant d'indiquer à une synthèse vocale quand elle doit lire ou épeler un sigle, avec cette fois un *ou* exclusif, car nous n'introduisons pas de variabilité : quand des sigles ont 2 réalisations, nous n'avons pas cherché laquelle était la plus fréquente, mais retenu celle qui obéissait à une règle plus générale. Or les conditions suivantes ne sont pas suffisantes.

1.2. LE GROUPE CV

Pour qu'un sigle soit lu, il faut qu'il comprenne un groupe CV (C=consonne, V=voyelle), lequel constitue la syllabe la plus naturelle [Rodríguez 80] : une consonne et une voyelle immédiatement subséquente, en effet, ne sont jamais séparées par une coupe syllabique [Wioland 81].

Nota : c'est à cette unité CV que nous faisons appel pour nommer les consonnes occlusives, qui ont besoin plus que toutes autres d'être soutenues par des voyelles, pour être produites et perçues - on les appelle /be/, /de/, etc, bien que ce soit redondant du point de vue de la théorie de l'Information [Chao 68] -; mais nous parlons ici des lettres et non de leurs prononciations.

Cela est en accord avec les principes de découpage syllabique [Gardes-Tamine 88]. La syllabe française est vocalique - on compte autant de syllabes qu'il y a de phonèmes vocaliques (les sonantes ne peuvent en être le sommet) -, et la fréquence des syllabes ouvertes est nettement supérieure à celle des syllabes fermées - plus généralement, celle du type CV à celle de tous les autres [Arrivé et al. 86]. Ce trait est commun aux langues romanes, mais pour une telle préférence, les faits suivants placent le français en première position, devant l'italien et même l'espagnol : le /s/ comme marque de pluriel s'est perdu et les affriquées ont été réduites au cours de l'histoire [Fouché 73].

Delattre a établi, à partir d'un corpus de phrases parlées, que 54,9 % des syllabes ont la structure CV. La terminologie confirme la prédominance de ce groupe - on dit qu'une voyelle est entravée lorsqu'elle ne termine pas une syllabe; certains parlent respectivement de *semi-consonne* ou de *semi-voyelle* selon qu'une glide commence ou termine une syllabe [Mounin 74]. Si l'on ne s'arrête pas à la frontière des mots, la liaison, l'enchaînement, l'élision, l'emploi de "cet" pour "ce", "mon" pour "ma" devant voyelle, etc y participent également [Kaye 84].

Ces phénomènes évitent l'hiatus. Il demeure que les sigles de la forme même CV, d'ordinaire, sont épelés - ce qui multiplie les hiatus (notamment /e/+voyelle) si

C ∈ {'B','C','D','G','J','K','P','Q','T','V','W'}

(ex. "BO", "KO", "JO", "QI").

Inversement, "ARC" est lu, alors qu'il ne présente pourtant pas de séquence CV. La condition susdite n'est donc ni nécessaire ni suffisante.

1.3. LA MORE

Citons à présent [Plénat 93] :

"C'est pour rendre compte de cette absence de correspondance qu'a été créée la notion de poids ou more. Une syllabe compte au maximum pour 2 mores; sauf en cas de synalèphe devant une autre voyelle, une voyelle compte pour une more, la ou le(s) consonne(s) finale(s) compte(nt) aussi, s'il y en a une ou plusieurs pour une more; les consonnes d'attaque n'interviennent pas dans le calcul. Si l'on mesure les sigles à cette aune, on dira que, pour être lu, un sigle doit avoir une réalisation qui 'pèse' au moins 2 mores."

Cette condition n'est pas suffisante : on épelle toujours "API".

1.4. LE PHONEME

M. Plénat ajoute, dans un article plus récent [Plénat 93] :

"On serait tenté de dire, sans plus, que pour qu'un sigle soit lu, il faut que sa lecture aboutisse à une forme comprenant au moins 3 phonèmes."

Cette règle explique bien pourquoi

- "IO" (p. ex.) se dit /io/ et non /jo/,
- des sigles de 3 lettres, bien que présentant une alternance de voyelle et de consonne, mais finissant par un *E* ou comprenant un *H*, sont épelés (ex. "ALE", "HEC").

Son auteur reconnaît que cela ne s'applique pas à "LU" (la fameuse marque de biscuits). Mais surtout, cette condition, comme les 2 précédentes, n'est pas suffisante pour qu'un sigle soit lu : en français, "ETA", "ITA" et "UTA" sont épelés. "Vérité en deçà des Pyrénées, erreur au-delà"? - l'acronyme basque pour "Euzkadi ta Askatasuna" est oralisé /eta/ du côté espagnol.

Le genre du sigle (qui est sauf exception celui de la tête), le domaine d'appartenance (ex. enseignement), le fait que la source soit syllabique et non alphabétique, le nom des lettres (on dit /ve/, mais /ef/) ne semblent pas non plus être déterminants.

2. CHOIX DU CORPUS ET ANALYSE

Le corpus de base que nous avons utilisé est celui de Marc Plénat : un millier de sigles avec le nombre et la répartition des lettres qui entrent dans leur composition, leur signification et leur(s) oralisation(s) - 2 en sont données, par exemple, pour "US", qui signifie à la fois "United States" et "L'Université Syndicaliste". Il se présente sous la forme d'un fichier SIGLES.DBF (logiciel de banque de données DBASE III). Nous avons à partir de lui établi 3 niveaux de règles, dont le premier est le plus général.

Notre analyse repose sur le nombre de lettres constituant le sigle. Ce critère est purement typographique, mais qu'est-ce qu'un sigle, sinon un jeu sur l'écriture alphabétique? Il n'est pas non plus aussi éloigné de la phonémisation qu'il peut le sembler au premier abord, car le sigle est avant tout une abréviation écrite, un langage visuel, une économie de typographie. "CNI", par exemple, est tout autant syllabable que "CNIL"; or il est épelé, alors qu'il serait plus court de le lire - les noms des consonnes graphiques ('B', 'C', etc) sont réalisés avec au moins 2 phonèmes.

Le fait que dans "PC", le 'C' (/se/) puisse avoir pour sources les phonèmes /s/ ("ceinture"), /k/ ("communiste") et /ʃ/ ("chimie") renforce l'idée que le sigle est essentiellement une création de la langue écrite. C'est cette dernière qui, à l'origine de telles modifications phoniques, prend le pas sur l'oral et se montre pour une fois plus dynamique [Calvet 70].

Notre programme commence donc par évaluer la longueur des sigles. Ceux-ci comportant de 2 à 9 lettres (maximum rarement atteint), on peut penser que l'homme fait de même, pour les oraliser : l'oeil humain peut les visualiser dans leur globalité.

2.1. REGLES DE NIVEAU 1

Le principe d'économie linguistique (application à la fonction de communication de la loi du moindre effort - limiter l'énergie dépensée tout en maintenant l'intercompréhension [Martinet 67] -) ne vient forcer celui de syllabicité qu'à partir de 4 lettres : au delà de 3 lettres (ce qui représente les sigles les plus fréquents), se pose en outre un problème mémoriel - les sigles sont plus difficiles à retenir. Ainsi :

- Les sigles de 2 lettres sont presque toujours épelés ("IA", "AG", "CD", "CA").
- Les sigles de 3 lettres sont majoritairement lus, s'ils sont de la forme CVC ou VCV; sinon, ils sont majoritairement épelés.
- Les sigles de 4 lettres de la forme CVCV, VCVC, CVVC, VCCV, VVCV et VCVV sont le plus souvent lus; sinon, ils seraient trop longs à épeler.
- Il en serait de même des sigles d'au moins 5 lettres, qui sont donc généralement lus, tant qu'ils contiennent une voyelle.

Les mots "cégétiste" (apparu dès le début du siècle), "(h)achélémisation", "ovniologie", "sidéen" (voire "sidaïque") et l'écriture de "sida" - qui perd même sa première majuscule, contrairement à "Onu", qui est un nom propre -, le créole "rémi" pour "RMI" et des acronymes entrés dans la langue, comme "laser" ou "radar", confirment ces faits.

2.2. REGLES DE NIVEAU 2

D'après ce que nous avons dit ci-dessus, nous prenons donc le parti d'épeler les sigles de 3 lettres, si la 1ère est d'un groupe différent de celui de la 2ème et du même groupe que la 3ème.

N.B. : on définit ici par *groupe* la classe des voyelles ou des consonnes graphémiques.

Cependant, il faut ajouter pour être plus précis que de tels sigles sont épelés

- s'ils comprennent un 'H', (ex "HIV").
- si la 2ème ou la 3ème lettre est un 'E' (ex. "CES").- si la 2ème lettre est une consonne momentanée - graphèmes 'B', 'C', 'D', 'G', 'P', 'T' - (ex. "OPA").

Cette opposition entre consonnes momentanées et continues intervient également dans les sigles de 4 lettres : ceux en C_1C_2VC et CVC_3C_4 sont généralement lus, si C_1, C_2 ou C_3 est une consonne continue : ex. "SMIC" (voir "smicard").

Quant au *E*, sa présence en fin de mot entraîne généralement l'épellation des sigles de la forme VCC*E* (ex. "OCDE"). Son élision, en effet, ramène ces derniers à la forme VCC, qui, on l'a vu, aboutit à une prononciation alphabétique (cf 2.1.).

2.3. REGLE DE NIVEAU 3

Autre règle relative au 'E' : les sigles en C_1'E'C_2 sont d'ordinaire lus si C_2='P' (ex. "NEP".

"PEP", "ZEP" - où l'analogie avec "ZUP" peut également jouer -).

2.4. EXCEPTIONS

Nous avons isolé les sigles qui, d'après le corpus, n'obéissaient pas à ces règles. Ils sont au nombre de 51 - nous n'avons pas retenu "CAP" ni "SEPT", dont nous connaissions une prononciation différente, allant avec une acception différente de ce que donne SIGLES.DBF - : nous les avons insérés dans de courtes phrases et avons fait lire à haute voix ce nouveau corpus à 15 personnes - en les enregistrant -, après leur avoir demandé de lire le texte une première fois pour elles-mêmes. Nous ne prétendons pas que cet échantillon - de 9 hommes et 6 femmes, de 19 à 85 ans - soit représentatif de la population française; nous avons toutefois entrepris cette démarche pou juger du caractère particulier de ces sigles faisant exception.

14 sigles, dont "PRI", par exemple, sont lus selon SIGLES.DBF - c'est en effet la prononciation espagnole pour ce parti politique mexicain -, alors que (à l'exception de 4 d'entre eux), ils ont tous été épelés à une forte majorité par nos informateurs, en accord avec nos règles. "ANGE", "CLE", "ARC" et "LU" (dont nous avons déjà parlé) aboutissent majoritairement, dans notre enquête, à une prononciation lue, selon un processus de jeu de mot; mais nous ne retiendrons comme exception que ces 2 derniers, qu'un seul locuteur a épelés.

Nous ne ferons de même, dans les sigles épelés dans SIGLES.DBF et lus de par nos règles, que lorsqu'ils sont épelés à l'unanimité (ex. "UEFA") ou à l'unanimité moins 1 (ex. "SIVP"), dans notre enquête. Tel est le cas de 24 sigles : l'épellation a été consacrée par l'usage dans "VIP" et "DUT" (où l'analogie avec "IUT", également, tient certainement une grande part), elle évite de plus une coïncidence malheureuse dans "USA", "CIC", "COD", "SAS" et "RAS"; elle s'explique sinon par la présence de séquences inhabituelles (ex. 'G"F', 'C"V'), ou encore d'une suite de 3 voyelles ou 3 consonnes dans des sigles qui comptent pourtant plus de 4 lettres (ex. "FNSEA"). Ce dernier résultat est conforme à la structure syllabique du français : le groupe CCC se rencontrant uniquement sous la forme

fricative+occlusive+liquide

(ex. "arbre"), il ne fait partie que des types de syllabes les moins courants [Wioland 85].

Les 13 sigles restants, parmi ceux qui échappent à nos règles, sont tous épelés d'après SIGLES.DBF. Chacun l'a été par au moins 8 personnes de notre échantillon, mais a aussi été lu par au moins 2 personnes. Outre le fait que ces sigles sont moins connus que les précédents (comme le confirment les commentaires de quelques informateurs), nous n'avons donc pas cru bon de les faire figurer comme exceptions. "BOF" sera ainsi lu par la synthèse - comme il l'est d'ailleurs par la majorité des moins de 30 ans, alors qu'on disait /beoef/, pendant la guerre, pour désigner les vendeurs au noir de "beurre, oeufs, fromage". Il en sera de même pour "CAT", l'hésitation marquée par un des locuteurs (qui l'a lu avant de se reprendre) illustrant bien que les 2 oralisations sont possibles.

3. PARTICULARITÉS

Nous avons parlé jusqu'ici de sigles lus ou épelés comme si, une fois cette partition faite, le problème de leur oralisation était résolu.

3.1. SIGLES ÉPELÉS

Pour la prononciation par lettres séparées, en effet, il n'y a guère le choix. Certes le "Toulouse Football Club" s'appelle /tefese/; 'E' peut être oralisé /ə/ ou /e/ (ex. "EOR") et 'W' comme 'V' (ex. "WC"). "GI", "DJ", "CB" sont parfois épelés à l'anglo-américaine (d'où "cibiste"). Et la tendance à dire "2E" pour "EE", "3A" pour "AAA", etc semble se répandre, même si on ne la rencontre dans SIGLES.DBF qu'avec "CII".

Mais surtout, la lecture (ou prononciation liée) des sigles diffère en plusieurs points de la lecture normale.

3.2. SIGLES LUS

Les consonnes finales sont toujours prononcées, ce qui n'est pas fait par VoxBox pour *-s* (même doublé), *-d*, *-t*, *-x*, *-r* : ex. "DDASS", "CAPES" (voir "capessien"). Sauf exception (dont "OTAN", pour laquelle il y a hésitation), les groupes 'AN', 'AM', 'EM', etc ne sont pas nasalisés - en fin de mot (ex. "GAN", "CNAM", "SACEM"), mais en début et en milieu de mot non plus (ex. "LIMSI"). Le 'E' est toujours prononcé /e/ en syllabe ouverte des sigles lus, même quand il est mis pour "emploi" : dans "ASSEDIC", notamment, il serait rendu caduc par VoxBox, si nous n'intervenions pas. Sans refaire toute la syllabation, notre programme regarde donc si le 'E' est suivi d'au moins 2 lettres, si la 2ème est un 'L', un 'R' ou une voyelle autre que 'E' final. Cela est suffisant pour les sigles : ainsi, en ne représentant que la transcription du 'E' dans le terme qui est à droite de la flèche

"ESA"	--> /e/
"SOGEGRAF"	--> /e/

mais

"ECTI"	--> /ɛ/
"ORSEC"	--> /ɛ/

Il faut également veiller en particulier aux sigles en *EN-* (signifiant souvent "École Nationale" - ex. "ENAC" -), pour que la synthèse ne prononce pas /ɑ̃/, comme dans "enivrer". Enfin, les groupes *-AI-* (ex. "MAIF") et *-EI-* ("SEITA") ne doivent jamais être phonémisés /ɛ/.

3.3. NOUVELLES EXCEPTIONS

Après avoir déterminé qu'un sigle doit être lu, nous avons alors traité ces particularités. Mais malgré ces règles, 27 nouveaux sigles sont prononcés autrement que dans SIGLES.DBF. Nous avons également relevé dans la presse audio-visuelle et dans la vie courante 18 sigles qui pouvaient échapper à nos règles (ex. "SSII", "MIT"). Comme ci-dessus, nous les avons soumis à 15 personnes (9 hommes et 6 femmes, de 19 à 85

ans), et n'avons retenu comme exceptions que les prononciations qui divergent pour au moins 13 locuteurs de ce que donnent nos règles. Ainsi, nous n'avons pas inclus dans notre programme les réserves faites en 3.1., qui concernaient 12 sigles de SIGLES.DBF.

En revanche, il nous a fallu prendre en considération "ASCII" (/aski/), "SDECE" (/zdek/), "CNASEA" (où le /S/ est dévoisé) et 4 sigles qui sont exceptionnellement oralisés avec une voyelle nasale, même si ces prononciations n'ont pas obtenu l'unanimité (même moins une voix). Le critère que nous avons suivi, en effet, est désormais moins strict, puisqu'il n'y a plus uniquement une alternative : il ne permet pas de déterminer la phonémisation de "SEEPU", qui a été épelé par tous nos locuteurs, alors que SIGLES.DBF en donne la transcription /sepy/; nous ferons dès lors en sorte que ce sigle se comporte - de même que "PEEP" et contrairement à "INSEE" - comme s'il ne comptait qu'un 'E'.

Le reste de notre enquête confirme le caractère exceptionnel de seulement 5 autres sigles, ne provenant pas de SIGLES.DBF : "ESPRIT", dont le *-t* final n'est bien sûr pas prononcé et 4 sigles, comme "SOS", qui sont épelés en dépit des règles générales. Nous aurions sans doute dû tenir compte de "GATT", de "AEAGC" et de sigles américains tels que "FBI", "CNN" ou "MTV", si nous les avions inclus dans l'enquête. Mais nous ne l'avons pas fait, et ils n'appartiennent pas à SIGLES.DBF. En résumé, sur plus d'un millier de sigles, nous n'avons observé - et traité - qu'une quarantaine d'exceptions aux règles indiquées en 2.1., 3.2. et 3.4.

4. RÉALISATION

4.1. MOYENS MIS EN OEUVRE

Nous avons conçu en Prolog un préprocesseur qui transforme le fichier à lire, de manière que les sigles soient phonémisés correctement : avant de soumettre un texte au synthétiseur, on active l'épellation des mots en majuscules - possibilité offerte par VoxBox - et le programme met en minuscules et traite les sigles qui doivent être lus (en prenant garde aux mots mêlant majuscules et minuscules). Le problème, que nous n'avons pas résolu, vient évidemment des mots écrits en majuscules (notamment les titres et les noms propres), qui sont dès lors traités comme des sigles.

Pour isoler les sigles d'éventuels signes de ponctuation

(, ; : ")

et ignorer une éventuelle apostrophe ainsi que ce qui la précèderait, la méthode de segmentation (ou "tokenisation") que nous avons adoptée consiste à faire un filtrage dès la lecture : retenir les positions des caractères tels que

(ESPACE TAB - ; ' " « »)

afin de les réinsérer dans la liste traitée. P. ex. :

"L'IRA --> IRA --> ira --> "L'ira
| | |
tokenisation traitement reconcatétion

Un *sigle* est dès lors détecté comme mot commençant par 2 majuscules. Parmi ces majuscules - "caps" -, on énumère les voyelles et les consonnes momentanées (moins nombreuses), et on définit les autres comme consonnes continues. On applique alors les règles citées en 2., on en déduit s'il faut ou non le convertir en minuscules

(code ASCII <-- code ASCII + 32)

et on le traite d'après 3.

4.2. RÉSULTATS

Nous avons testé notre réalisation sur un corpus réel d'environ 5 pages : un cours de télématique particulièrement riche en mots écrits en majuscules.

Le tableau ci-dessous compare la quantité de sigles lus et épelés, avant et après traitement, avec ce qu'on attend - quand nous parlons de *traitement* ici, il faut inclure l'activation de l'épellation des majuscules (cf 4.1.).

Sigles	lus	épelés
Avant traitement	14	1
Après traitement	4	11
Ce qu'on attend	6	9

La différence entre la 1ère ligne (où un unique sigle, sans voyelle, est épelé) et la 2ème ligne est donc sensible : notamment dans l'exemple de "TCP/IP" et "USA". Celle entre la 2ème et la 3ème vient seulement

- du chiffre romain "IV" (que nous avons rangé ici parmi les sigles, pour simplifier, et dont on souhaite l'oralisation /katR/),
- de "IBM", qui, collé à un "*d* apostrophe", n'est pas détecté comme mot en majuscules par VoxBox.

Cela souligne quelques lacunes dans les choix de tokenisation faits pour cette synthèse vocale - ce dont il faudra tenir compte si l'on veut mettre en application ces travaux. Mais dans les 2 cas, la synthèse ne donnait pas de meilleur résultat avant traitement - elle lisait en effet /iv/. De plus, "ENSI" est maintenant lu conformément à 3.2.

Le tableau suivant décrit le sort des mots en majuscules autres que les sigles :

Autres mots	lus	épelés
Avant traitement	51	0
Après traitement	46	5
Ce qu'on attend	51	0

Les 1ère et 3ème lignes sont évidemment les mêmes. Les mots injustement épelés après

traitement sont tous des mots-outils de 2 lettres : "ET" (apparaissant 2 fois), "LE", "LA" et "DE".

10 mots qui étaient bien lus par VoxBox ne le sont plus après traitement : ils sont encore lus, mais de la façon précisée en 3.2. Ces spécifications déplacent l'erreur dans "RESEAUX", 2 fois présent - le premier 'E' est bien phonémisé /e/, mais le 'X' est prononcé. Cependant elles rapprochent 6 autres mots - soit 9 occurrences - de leur prononciation anglaise (ex. "WORD").

CONCLUSION

Ainsi, la phonémisation du français est difficile jusque dans la siglaison. Nous nous sommes consacré à l'étude de ce point, qui est jusqu'ici resté marginale en phonologie.

Après avoir présenté les définitions générales et les théories actuelles, nous avons exposé 3 niveaux de règles, reposant sur le nombre de lettres qui entrent dans la composition des sigles. Puis nous avons étudié les particularités de l'épellation et de la lecture, à partir d'une base de données d'un bon millier d'acronymes. Les essais sur un corpus réel ont montré une amélioration sensible.

Notre objectif était d'augmenter la qualité de la synthèse vocale, la rendre plus intelligible et plus correcte, tout en nous limitant à la conversion graphème-phonème des sigles. Même si nous avons compté sur l'acquis de VoxBox dans ce travail, la plupart des idées que nous avons développées sont indépendantes du système utilisé. Or il n'est pas besoin ici de dire tout l'intérêt de la Synthèse Vocale.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Arrivé et al. 86] A. Arrivé, F. Gadet & M. Galmiche, 1986, *La grammaire d'aujourd'hui, guide alphabétique de linguistique française*, Flamarion, Paris.

[Béchade 92] H.-D. Béchade, 1992, *Phonétique et morphologie du français moderne et contemporain*, Presses Universitaires de France, Paris.

[Calvet 70] L.-J. Calvet, 1970, *Le phénomène des sigles en français contemporain* (thèse inédite de 3ème cycle), Université de Paris (Sorbonne).

[Carlson et al. 82] R. Carlson, B. Granström & S. Hunnicutt, "A multi-language text-to-speech module", *Proc. IEEE-ICASSP, vol. 3*, pp 1604-1607, Paris.

[Chao 68] Y. Ren Chao, 1968, *Langage et systèmes symboliques*, Payot, Paris.

[Doppagne 91] A. Doppagne, 1991, *Majuscules, abréviations, symboles et sigles. Pour une toilette parfaite du texte*, Ducilot, collection l'esprit des mots, Paris.

[Fouché 73] P. Fouché, 1973, *Phonétique historique du français, introduction*, Éditions Klincksieck, Paris.

[Gardes-Tamine 88] J. Gardes-Tamine, 1988, *La grammaire*, tome 1 : "Phonologie, morphologie, lexicologie", collection cursus, Armand Colin, Paris.

[Kaye 84] J.D. Kaye & J. Lowenstamm, 1984, "De la syllabicité", in F. Dell, D. Hirst & J.-R. Vergnaud (eds.), *Forme sonore du Langage. Structure des représentations en phonologie*, pp 123-159, Hermann, Paris.

[Martinet 67] A. Martinet, 1967, *Éléments de linguistique générale*, A. Colin, Paris.

[Mounin 74] G. Mounin, 1974, *Dictionnaire de la linguistique*, Presses Universitaires de France, Paris.

[Plénat 91] M. Plénat, 1991, "Vers une phonémisation des sigles", VRC-Greco.

[Plénat 93] M. Plénat, 1993, "Observations sur le mot minimal français, l'oralisation des sigles", à paraître dans B. Laks & M. Plénat, *De Natura Sonorum*, PUV, Paris.

[Rodríguez 82] F. Rodríguez González, 1982, "Variaciones fonotácticas en siglas : condicionamientos lingüísticos y sociolingüísticos", *Revista española de lingüística*, tome 12 vol. n° 2, pp 357-374.

[Wioland 85] F. Wioland, 1985, *Les structures syllabiques du français*, Slatkne-Champion, Genève-Paris.

[Wioland 91] F. Wioland, 1991, *Prononcer les mots du français. Des sons et des rythmes*, Hachette, Paris.

FORMALISATION DE CONNAISSANCES PROSODIQUES À L'AIDE D'UN SYSTÈME D'APPRENTISSAGE SYMBOLIQUE ET NUMÉRIQUE

Jean-Jacques SCHNEIDER , Pierre ROBINEAU, Henri MÉLONI

LABORATOIRE D'INFORMATIQUE D'AVIGNON
33, RUE LOUIS PASTEUR - 84000 AVIGNON - FRANCE

Résumé

Dans cet article nous présentons un système d'apprentissage symbolique et numérique basé sur la recherche de similarités (Similarity Based Learning) pour la formalisation de connaissances prosodiques sous forme de règles explicites.
Après une phase de classification supervisée des données numériques, le système détermine par généralisations l'ensemble des règles caractérisant les phénomènes observables sur le corpus.

Mots-clés :
prosodie, apprentissage, classification.

1. INTRODUCTION

L'utilisation de techniques d'apprentissage automatique en prosodie a déjà fait le l'objet de travaux, que ce soit par une approche connexioniste [Traber 92], ou une approche orientée connaissances [Mortamet & al. 90]. C'est cette dernière que nous avons retenue, car dans notre optique, un système d'apprentissage ne peut être validé que s'il est capable de présenter les connaissances acquises ainsi que d'expliquer les mécanismes mis en jeu.
L'utilisation conjointe de l'apprentissage numérique et symbolique dans la traitement automatique de la parole [Desté 93], ainsi que l'explication pour la validation de systèmes d'apprentissage [Nédellec 93], sont des approches abordées depuis peu.
A cause de la multiplicité des paramètres acoustiques mis en jeu et des différents niveaux d'analyse, il est primordial que le système ne dispose d'aucune connaissance prosodique préalable. De plus, pour chacun des paramètres, plusieurs descriptions peuvent être envisagées. Pour ces raisons nous avons choisi un étiquetage explicite sous une forme attribut/valeur (les valeurs pouvant être de type symbolique ou numérique). Le corpus d'apprentissage est constitué d'un ensemble de descriptions de l'énoncé sous la forme d'informations diverses (syntaxique, sémantique, phonétique, ...) et des paramètres prosodiques correspondants (durée, fréquence, énergie, ...).
La classification offre à l'expert la possibilité d'observer le comportement des données, pour une reformulation éventuelle des exemples (regroupements inattendus, regroupements attendus inexistants, représentation plus concise et structurée des informations, ...).
Notre système est volontairement épuré des heuristiques habituellement mises en place. Il est en effet conçu pour être utilisé par un expert en prosodie, et la réduction de l'espace de recherche ne peut dépendre que de la volonté de ce dernier et non de celle de l'apprentiste.

2. REPRÉSENTATION DES EXEMPLES

2.1. LANGAGE DE DESCRIPTION

Un exemple est constitué de deux parties : un ensemble de descriptions de paramètres que l'expert suppose être à l'origine du phénomène, et la description du phénomène lui-même.
Chacune des parties se décompose en ensembles de descriptions élémentaires attribut/valeur. Cette syntaxe a pour principal avantage de rendre les descriptions explicites. **Attribut** est une étiquette symbolique dont le sens est connu de l'expert. Le langage de généralisation (syntaxe des règles produites par le système) est déduit du langage de représentation des exemples, il hérite donc du caractère explicite de celui-ci. **Valeur** appartient à l'ensemble de définition des données du système.

2.2. LES DONNÉES DU SYSTÈME

Les deux types de base que sont le **numérique** et le **symbolique** permettent une formulation explicite et naturelle des données. Le premier permet la représentation d'informations sous une forme brute, sans interprétation. Le second permet la représentation de concepts abstraits à l'aide de symboles (chiffres, lettres, mots) dont la sémantique est connue. Il est impossible, avec les deux types précédents, de représenter un concept à plusieurs dimensions. Pour cela, il est nécessaire de définir un troisième type : le type **énuméré**. De telles données sont exprimées sous la forme d'une liste de valeurs numériques ou symboliques.

A la manière de [Michalski 83], les données sont regroupées en trois classes distinctes, les données atomiques, les données linéaires et les données structurées.
Les données atomiques ont des valeurs qui sont indépendantes les unes des autres. Par exemple, le mode articulatoire d'une voyelle peut être décrit par "ouverte", "fermée", "nasale" ou "orale".
Les données linéaires ont des valeurs dont le domaine de définition est un ensemble ordonné. Pour cette classe de données, il est nécessaire de définir une *relation d'ordre* et une *distance*. Dans le cas de données de type numérique, il peut s'agir de l'ordre naturel des nombres, et de la distance euclidienne. Pour les valeurs symboliques, cette notion devra être connue par avance. Par exemple, si la durée est décrite par les valeurs symboliques "très réduite", "réduite", "normale", "allongée", "très allongée", il est possible de reformuler ce paramètre en ["très réduite"..."réduite], "normale", ["allongée..."très allongée"].
Il existe une troisième classe, **les données structurées**, mais actuellement, aucun des paramètres retenus n'en nécessite l'usage.

Chaque attribut est associé à un type basique/énuméré, numérique/symbolique, à une classe, atomique/linéaire/structurée, et à un ensemble de fonctions permettant la manipulation de telles données.

3. DESCRIPTION DU CORPUS D'APPRENTISSAGE

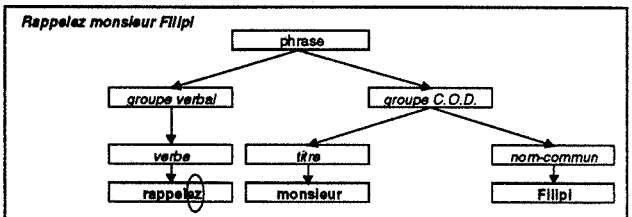
Pour mettre en pratique la syntaxe du langage de représentation des exemples définie dans la section précédente, nous présentons un étiquetage prosodique des énoncés.
Il est nécessaire de définir le cadre du phénomène à caractériser, c'est-à-dire l'unité linguistique minimale permettant la description d'un phénomène prosodique (voyelle, syllabe, mot, groupe de sens, syntagme, ...). Notre choix s'est porté sur l'étude des voyelles.
La première partie des exemples est constituée d'un ensemble d'informations diverses (syntaxe, contexte phonétique, ...), que nous supposons être à l'origine des phénomènes prosodiques pour le type de corpus étudié. Elle est appelée *description de l'énoncé*. La seconde partie, appelée *description des paramètres prosodiques*, est, comme son nom l'indique, la description de l'ensemble des paramètres prosodiques considérés.

3.1. DESCRIPTION DE L'ÉNONCÉ

Les descripteurs proposés sont, pour la plupart, issus des travaux de [G. Caelen-Haumont 91], et exploitent les possibilités du codage des données défini dans la section précédente.

Description de la syntaxe

A partir de la structure syntaxique d'un énoncé, il est possible d'extraire deux informations de niveaux différents.
La première information est purement de nature grammaticale. Il s'agit de la structure de l'énoncé, mis sous la forme d'une liste contenant les caractéristiques grammaticales du mot, du groupe. Pour ce descripteur, nommé *structure grammaticale*, chaque description élémentaire a une valeur de type basique symbolique et de classe atomique.
Pour faire apparaître les regroupements en unités grammaticales, ainsi que l'influence de la taille de celles-ci sur les paramètres prosodiques, le descripteur *structure syntaxique* contient, sous forme de liste, la place de chaque constituant dans le constituant de niveau supérieur. Cette position est donnée par le rang de l'élément et le nombre total d'éléments. C'est une valeur de type énuméré numérique et de classe atomique.
Voici une illustration des deux descripteurs précédents sur la phrase suivante, en considérant la voyelle [e] :




```
(structure_grammaticale
    ((sg_mot "verbe") (sg_groupe "groupe verbal")))
(structure_syntaxe
    ((ss_voyelle (3 3)) (ss_mot (1 1)) (ss_groupe (1 2))))
```

Description du contexte phonétique

Ce descripteur permet de mettre en évidence l'influence des modes articulatoires de la voyelle considérée, mais également celle de son contexte phonétique. Ce dernier est constitué des modes articulatoires de la consonne précédente et de la consonne suivante. Les trois descripteurs précédents sont constitués de descripteurs élémentaires ayant des valeurs de type basique symbolique et de classe atomique. Un quatrième descripteur permet l'observation des effets de la position de la pause sur les paramètres prosodiques. C'est une valeur de type basique numérique et de classe linéaire.
Voici une illustration avec la phrase précédente, pour la voyelle [e].

```
(voyelle
    ((v_mode1 "orale") (v_mode2 "fermée")))
(consonne_prec
    ((cp_mode1 "liquide") (cp_mode2 "voisée")))
(consonne_suiv
    ((cs_mode1 "nasale") (cs_mode2 "voisée")))
(pause
    ((p_distance 10)))
```

Description de la sémantique

Ce descripteur ne fait pas partie de l'étiquetage retenu. La nature des énoncés du corpus ne se prête pas à une telle analyse. Citons deux descripteurs issus des travaux de [G. Caelen-Haumont 91] qui auraient pu figurer dans cette partie. Le premier spécifie la nature support/ apport de chaque constituant dans l'analyse syntaxique d'un énoncé, le second indique le registre (fondamental/spécialisé) et le référent (concret/abstrait) du syntagme contenant la voyelle. Les valeurs des descripteurs élémentaires pourraient être de type basique symbolique et de classe atomique.

3.2. DESCRIPTION DES PARAMÈTRES PROSODIQUES

Les paramètres prosodiques étudiés sont la durée, la fréquence fondamentale et l'énergie. Pour chacun de ces paramètres, un ou plusieurs descripteurs sont proposés.

Description de la durée

Connaissant la durée intrinsèque de chaque voyelle, pour un débit donné, il est possible de déterminer la différence entre la valeur attendue et la mesure effective. Cette différence est exprimée sous la forme d'un pourcentage. Les valeurs de ce descripteur sont de type basique numérique et de classe linéaire.
Voici deux exemples de description de la durée issus du corpus d'apprentissage.

```
(duree
    ((d_alteration 20.3)))
(duree
    ((d_alteration -26)))
```

Description de la fréquence fondamentale

Pour pouvoir observer des phénomènes globaux, et des phénomènes locaux, il est nécessaire de disposer de deux descripteurs pour la fréquence fondamentale. Le premier, appelé contour global, décrit l'évolution de la fréquence de la voyelle étudiée par rapport à celle des voyelles adjacentes. Le second appelé contour local, décrit l'évolution de la fréquence sur la voyelle étudiée.
Un contour est décrit par trois valeurs. Pour le contour global, il s'agit de la valeur de la fréquence au milieu de la voyelle courante, et au milieu des voyelles adjacentes, pour le contour local, des valeurs de la fréquence au premier tiers, au milieu, et au dernier tiers de la voyelle. Pour pouvoir caractériser les phénomènes contextuels, chaque contour est décomposé de deux descripteurs élémentaires : le premier décrit la partie gauche du contour, le second la partie droite. Les valeurs de ces descripteurs sont de type énuméré numérique et de classe linéaire.
Voici un exemple de descriptions de la fréquence fondamentale issu du corpus d'apprentissage.

```
(f0_melodique
    ((f0_g_av (137 137)) (f0_g_ap (137 159)))
(f0_micromelodique
    ((f0_l_av(167 137)) (f0_l_ap (137 132)))
```

Description de l'énergie

La valeur de l'énergie est normalisée par rapport à sa valeur moyenne sur l'énoncé. Comme pour la durée, chaque voyelle dispose d'une énergie intrinsèque. Cette dernière est exprimée sous la forme d'un pourcentage représentant la différence entre la mesure effective et la valeur attendue. C'est une valeur de type basique numérique et de classe linéaire.
Voici un exemple de description de la durée issu du corpus d'apprentissage.

```
(energie
    ((e_alteration 11.4)))
```

Le choix des descriptions pour l'énoncé et pour les paramètres prosodiques, ainsi que les valeurs choisies sont bien évidemment critiquables. le but n'est pas de définir un étiquetage universel, mais de présenter les possibilités de codage du système à travers un étiquetage adapté à notre corpus.

3.3. LE CORPUS D'APPRENTISSAGE

Le corpus de départ, constituée d'un ensemble de 48 phrases, est un sous-ensemble du corpus défini pour le projet Gréco-Prosodie. Les phrases sont prononcées par un même locuteur, à un débit moyen de 7 pho/s (phrases lentes).
De ce corpus sont extraits automatiquement 518 exemples dont la syntaxe est définie précédemment.
A titre d'illustration, voici un exemple appartenant au corpus d'apprentissage :

```
((description_enonce
    ((structure_syntaxique
        ((ss_voyelle (4 6)) (ss_mot (1 1)) (ss_groupe (1 4))))
     (structure_grammaticale
        ((sg_mot "verbe") (sg_groupe "groupe_verbal")))
     (voyelle
        ((v_mode1 "orale") (v_mode2 "fermee")))
     (consonne_precedente
        ((cp_mode1 "occlusive") (cp_mode2 "non_voisee")))
     (consonne_suivante
        ((cs_mode1 "liquide") (cs_mode2 "voisee")))
     (pause
        ((p_distance 12)))))
(description_parametres_prosodiques
    ((f0_micromelodique
        ((f0_l_avant (167 167)) (f0_l_apres (167 161))))
     (f0_melodique
        ((f0_g_avant (159 167)) (f0_g_apres (167 120))))
     (duree
        ((d_alteration 6)))
     (energie
        ((e_alteration 71))))))
```

4. LE SYSTÈME D'APPRENTISSAGE

4.1. PRINCIPE

Le système d'apprentissage est constituée de deux parties.
La classification des valeurs des paramètres prosodiques offre à l'expert, en plus de l'observation du comportement des données, la possibilité de déterminer une représentation des phénomènes. La généralisation sur la description de l'énoncé permet de déterminer les descripteurs significatifs, c'est-à-dire une caractérisation des phénomènes prosodiques observables sur le corpus.

4.2. LA CLASSIFICATION DES DONNÉES

Principe

L'ensemble des N exemples du corpus d'apprentissage peut être considéré comme un ensemble de N règles de la forme : (description de l'énoncé) ➡ (description des paramètres prosodiques).
Parmi cet ensemble de règles, il en existe de nombreuses dont les descriptions de l'énoncé (DE) sont strictement identiques. De telles règles sont regroupées en une seule règle dont chacun des descripteurs élémentaires (Delt) de la description des paramètres prosodiques (DPP) est une classification de l'ensemble des valeurs de chacun des Delt des règles de départ.
Cette étape de classification apporte une information qui se situe au niveau sémantique. Comme le souligne [Lance & Williams 67], "il ne s'agit pas d'atteindre un résultat vrai ou faux, probable ou improbable, mais seulement profitable ou non profitable". En effet, la classification offre à l'expert la possibilité d'observer le comportement des données : des regroupements inattendus apparaissent, de nouvelles hypothèses sur les données sont nécessaires ; des regroupements attendus n'existent pas, faisant ressortir le faible pouvoir séparateur des paramètres utilisés. Les classes ainsi obtenues offrent une représentation plus concise et structurée des données, en l'occurrence chacun des paramètres prosodiques.

Représentation des classifications

Une classification est composée d'un ensemble de classes, chacune étant une représentation possible du paramètre prosodique considéré. Une classe est définie par le nombre d'éléments qui la compose, ainsi que d'une valeur la décrivant. Plusieurs représentations sont envisageables pour cette dernière :
① Dans le cas de données de classe linéaire, et de type numérique, symbolique ou énuméré, la valeur d'une classe peut être sous la forme d'un intervalle. Pour le traitement des données de classe atomique, l'intervalle se réduit à un singleton. Voici un exemple illustrant cette représentation :

```
(duree (classif_intervalle (12 (-10 -5)) (5 (8 15)) (3 (58 62 ))))
```

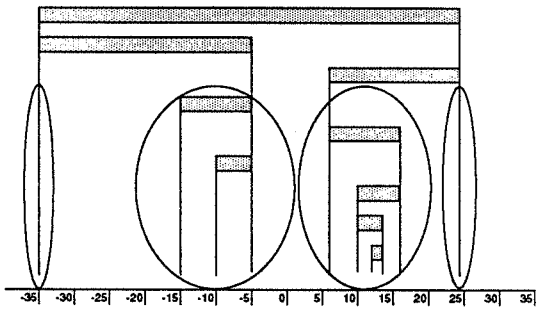
② Dans le cas de données de classe linéaire et de type numérique ou énuméré numérique, la valeur d'une classe peut être sous la forme d'une valeur moyenne. Pour le traitement des données de classe atomique ou de type symbolique, la valeur moyenne est la valeur de départ. Pour illustrer cette représentation, l'exemple suivant reprend les valeurs de l'exemple précédent :

```
(duree (classif_moyenne (12 -7) (5 11) (3 60)))
```

Les méthodes de classification

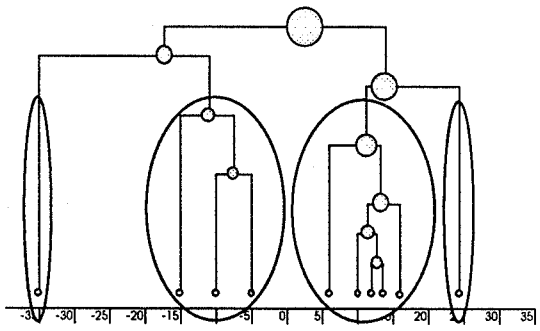
Parmi les méthodes de classification des données existantes, seules la classification hiérarchique et les nuées dynamiques sont actuellement implantées dans le système ([Ward 63], [Diday & al. 80], [Diday & al. 89])

La classification hiérarchique, grâce à une possible interprétation visuelle, est un outil très simple d'accès pour un individu non expert en traitement de données. A chaque classification, le système propose une hiérarchie sous forme de dendogramme. A l'expert de déterminer les classes nécessaires pour la description du phénomène qui lui est présenté.
Voici une illustration des deux classifications présentées, sous forme d'intervalle et de moyenne, à partir de l'ensemble de valeurs (-5 -35 +13 +24 -10 +10 -14 +16 +6 +12), l'expert choisissant quatre classes :



Voici la liste produite à l'aide d'une classification sous forme d'intervalle :

```
(classif_intervalle
    (5 (6 16)) (3 (-14 -5)) (1 (-35 -35)) (1 (24 24)))
```



Voici la liste produite à l'aide d'une classification sous forme de moyenne :

```
(classif_moyenne
```

```
(5 12) (3 -9) (1 -35) (1 24))
```

La méthode des **nuées dynamiques** est actuellement en cours d'implantation. Le peu de résultats obtenus, ne permettent pas de présenter ces travaux dans cet article.

4.3. LA GÉNÉRALISATION

Le principe de base est de rechercher, parmi l'ensemble des n descripteurs élémentaires de la DE, ceux qui ont une réelle influence sur la DPP.

Pour cela, la généralisation fonctionne de manière itérative à partir de l'ensemble des N règles R_n obtenues précédemment. A chaque étape i, de 1 à n-1, les p règles qui ont des DE identiques, à i descripteurs près, sont isolées de l'ensemble R_{n-i+1}.
Trois cas peuvent alors se présenter :
① Tous les descripteurs ont un *comportement comparable*. Les p règles sont supprimées de l'ensemble R_{n-i+1}, et une règle, dont les i descripteurs de la DE sont associés à une valeur prédéfinie, est ajoutée à l'ensemble R_{n-i}.
② Uniquement certains descripteurs ont un *comportement comparable*. Une nouvelle règle est ajoutée à l'ensemble R_{n-i}, dont les descripteurs de la DPP qui n'ont pas un tel comportement sont associés à une valeur prédéfinie. Contrairement au cas précédent, les p règles ne sont pas supprimées de R_{n-i+1}, mais les valeurs des descripteurs de la DE qui ont un *comportement comparable* sont associés à une valeur prédéfinie.
③ Aucun descripteur n'a un *comportement comparable*, l'ensemble R_{n-i+1} est laissé tel quel.

La notion de *comportement comparable* implique la définition d'une comparaison entre classifications.
Dans le cas de la classification sous forme d'intervalle, il est possible de comparer n classifications en observant, en collaboration avec l'expert, l'intersection de toutes les classes. Pour un meilleur résultat il est important de pondérer cette intersection par la taille des classes. Pour les classifications sous forme de moyenne, le principe est sensiblement le même.

5. RÉSULTATS ET CONCLUSIONS

La validation du système auprès des experts est en cours. Il produit déjà des règles qui semblent intéressantes. Voici un exemple dont le sens est : "en début de phrase, la première voyelle (structure syntaxique) d'un verbe (structure grammaticale) subit dans 69% des cas une altération de la durée de 6 à 16% (valeur de la plus grande classe du paramètre)".

```
((description_enonce
    ((structure_syntaxique
        ((ss_voyelle (1 4)) (ss_mot (1 1)) (ss_groupe (1 4))))
     (structure_grammaticale
        ((sg_mot "verbe") (sg_groupe "groupe verbal")))
     (voyelle
        ((v_mode1 !) (v_mode2 !)))
     (consonne_precedente
        ((cp_mode1 "vide") (cp_mode2 "vide")))
     (consonne_suivante
        ((cs_mode1 !) (cs_mode2 !)))
     (pause
        ((p_distance !)))))
(description_parametres_prosodiques
    ((f0_micromelodique
        ((f0_l_avant ?) (f0_l_apres ?)))
     (f0_melodique
        ((f0_g_avant ?) (f0_g_apres ?)))
     (duree
        ((d_alteration
          (classif_intervalle
            (27 (6 16)) (9 (-14 -5)) (2 (-35 -36)) (1 (24 24))) 39)))
     (energie
        ((e_alteration ?))))))
```

Chaque règle construite contient le nombre total de règles qui ont été généralisées pour l'obtenir. L'expert dispose donc d'une information lui permettant de juger de la pertinence de chacune des règles ainsi produite.
Malgré les défauts inhérents à la classification hiérarchique (sensibilité au bruit, à l'ordre des données, ...), le système se comporte de manière satisfaisante.
Néanmoins des problèmes subsistent lors de la comparaison de classifications. L'ouverture vers d'autre algorithmes de classification de données, automatiques ou supervisés, qu'il sera possible de choisir en fonction de la nature des informations à traiter, est envisagée pour la résolution de ce problème.

BIBLIOGRAPHIE

[Caelen 91]
G.Caelen-Haumont. *Stratégie des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions pragmatiques et paramètres prosodiques.* Thèse d'Etat - 1991.

[Diday & al. 80]
E. Diday et collaborateurs. *Optimisation en classification automatique.* INRIA.

[Diday & al. 89]
E.Diday G.Celeux G.Govaert Y.LecheValeurlier H.Ralambondrainy *Classification automatique des données - Environnement statistique et informatique.* Dunod informatique.

[Desté 93]
O.Desté. *LVQ+CHARADE = Un exemple d'apports mutuels entre apprentissages numérique et symbolique.* Journée Acquisition Validation Apprentissage - Saint Raphaël 1993.

[Michalski 83]
R.S.Michalski. *A theory and methodology of inductive learning.* Machine Learning Vol. 1 - Tioga Palo Alto 1983.

[Mortamet & al. 90]
L.Mortamet F.Emerard L.Miclet. *Attempting automatic prosodic knowledge acquisition using a database.* Workshop on speech synthesis - Autran 1990.

[Nédellec 93]
C. Nédellec. *Expliquer pour valider les systèmes d'apprentissage automatique.* Journée Acquisition Validation Apprentissage - Saint Raphaël 1993.

[Lance & Williams 67]
G.N.Lance W.T. Williams. *A general theory of classification sorting strategy, 1: hierarchical systems.* Computer Journal, Vol. 9 n°4.

[Schneider 92]
J.J.Schneider. *Mise en place d'un système d'apprentissage adapté à la prosodie du français.* Séminaire Prosodie - La Baume-lès-Aix 1992.

[Schneider 93]
J.J.Schneider. *Un système d'apprentissage pour la formalisation de connaissances. Application à la prosodie du français* Congrès Jeunes Chercheurs en Sciences Cognitives - La Motte d'Aveillans 1994 (à paraître).

[Traber 92]
C.Traber. *F0 generation with a database of natural f0 patterns and with a neural network.* Talking Machines : Theories, Models and Designs - North Holland 1992.

[Ward 63]
J.H.Ward. *Hierarchical grouping to optimize an objective function.* JASA Vol. 58.

CONTROLEUR GESTUEL DES CORDES VOCALES D'UN ROBOT PARLANT

VESCOVI Christophe, CASTELLI Eric

Institut de la Communication Parlée U.R.A. - CNRS N° 368
I.N.P.G./E.N.S.E.R.G. - Université STENDHAL
46, Avenue Félix Viallet, 38031 GRENOBLE Cedex 1

Résumé

Jusqu'à présent, les modèles d'inversion du signal de parole ne se sont intéressés principalement qu'à la représentation du conduit vocal. Notre étude étend ces modèles d'inversion (approche connexioniste) à la source vocale. Le modèle de source utilisé dans notre synthétiseur est le modèle à deux masses des cordes vocales. La caractérisation du débit glottique produit par ce modèle physique est réalisée dans un premier temps afin de définir un espace de contrôle approprié. Le contrôleur gestuel est alors doté d'un modèle direct (ou copie "mentale" ou efférente) du modèle à deux masses afin de permettre le passage entre les commandes articulatoires de la source et les paramètres de contrôle. Le contrôleur est validé en inversant des signaux de synthèse et en comparant les commandes estimées avec celles utilisées par le robot parlant. Des premiers résultats sur l'inversion d'un signal de parole naturelle sont finalement présentés.

Mots-clés : inversion, source vocale, modèle à deux masses, modèle LF, filtrage-inverse.

INTRODUCTION

Le modèle de simulation temporelle utilisé à l'I.C.P., appelé S.I.M.O.N.D. (Castelli & Badin 1993) peut se décomposer en trois parties comme le montre la figure n°1 : deux modèles acoustiques basés sur le principe de Kelly & Lochbaum (1962) pour les poumons et le conduit vocal et un modèle physique des cordes vocales, le modèle à deux masses (Ishizaka & Flanagan, 1972). Les interactions entre les cavités subglottiques et supraglottiques et entre le modèle mécanique de la glotte ont été particulièrement étudiées (Trinh V. et al. 1991). Afin de reproduire simplement les stratégies articulatoires désirées, le modèle de production de la parole a été couplé au modèle articulatoire de Maeda (1990). Les paramètres de commandes de l'ensemble complet sont alors réduits à onze : hauteur et protrusion des lèvres, apex, corps et dos de la langue, mâchoire, larynx et velum pour définir la géométrie du conduit vocal, auxquels il faut ajouter la pression subglottique (Ps) pour les poumons, l'ouverture des cordes vocales au repos (Ag0) et le facteur masse-tension (Q), paramètres de commande du modèle des cordes vocales. L'ensemble modèle articulatoire et modèle de production constitue ainsi le "plant" de notre robot parlant.

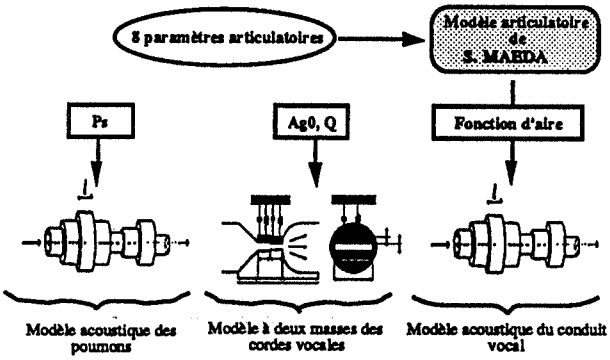


Figure n° 1 : Modèle de simulation temporel de l'ICP.

L'approche robotique sur laquelle ce travail s'appuie consiste à construire un contrôleur gestuel permettant d'apprendre la correspondance entre une séquence phonologique donnée et les trajectoires articulatoires devant être fournies au robot parlant. Ce problème est connu sous le terme général d'inversion. L'inversion du conduit vocal a déjà été réalisée à l'I.C.P. (Bailly et al., 1991) et l'approche connexioniste utilisée alors est conservée et étendue en introduisant la source vocale.

Pour corriger en permanence ses prédictions, le contrôleur (figure n° 2) doit se doter d'une copie "mentale" ou efférente du système à contrôler. Cette copie "mentale", appelée modèle direct, doit approximer la sortie *Y* qu'aurait eu le modèle articulatoire si on lui avait appliqué la commande *X*. Le contrôleur doit être doté d'un traitement perceptif adapté aux signaux glottiques. Alors que les commandes

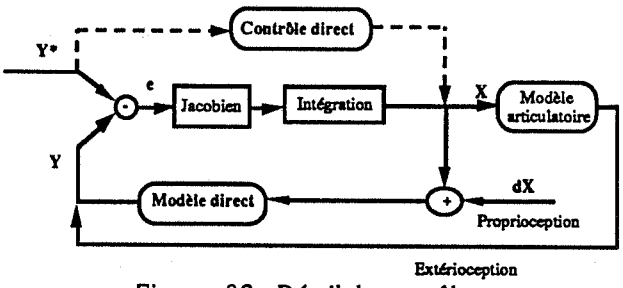


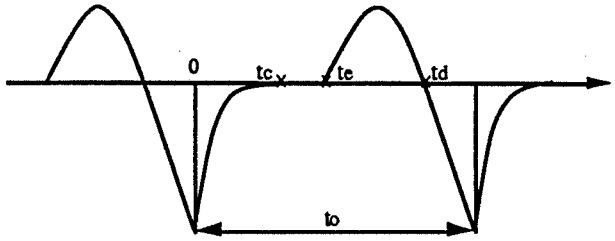
Figure n° 2 : Détail du contrôleur
Y est le point de l'espace distal que l'on veut atteindre. L'erreur e commise sur Y est corrigée grâce au modèle direct par descente de gradient. Une première approximation des commandes X est obtenue par contrôle direct, ce qui permet une stabilisation du contrôle.*

articulatoires/aérodynamiques sont celles du modèle à deux masses, l'espace de contrôle sera la forme du débit glottique produit par ce dernier.

CARACTERISATION DU DÉBIT GLOTTIQUE

Afin d'évaluer le signal d'excitation du conduit vocal mesuré sur du signal de parole naturelle, un outil de filtrage inverse a été mis au point. Cet outil est basé sur l'algorithme IAIF proposé par ALKU et al. (1991). Afin de rendre plus automatique les procédures de filtrage inverse, cet algorithme utilise quatre analyses LPC d'ordres différents. Une première analyse d'ordre 1 qui permet d'évaluer la pente du spectre du signal de parole, correspond à une préaccentuation adaptative. Une deuxième analyse d'ordre plus élevé (classiquement 12) permet d'estimer une première fois la fonction de transfert du conduit vocal : en éliminant par filtrage inverse la contribution du conduit vocal dans le signal de parole, on obtient une première estimation du débit glottique. La contribution de ce débit glottique va être ensuite estimée et éliminée du signal de parole grâce à une troisième analyse d'ordre 2 ou 4 avant d'évaluer une dernière fois la fonction de transfert du conduit vocal (quatrième analyse).

Le débit glottique ainsi obtenu, doit être caractérisé afin d'être reproduit. Pour diminuer la complexité du contrôleur gestuel cette caractérisation est réalisée par un nombre restreint de paramètres. En premier lieu, la dérivée d'une onde glottique stylisée, résultat d'un modèle analytique du débit glottique, est ajustée sur l'approximation de la dérivée du débit glottique obtenue par filtrage inverse. Le modèle analytique choisi est le modèle LF (Fant et al. 1985) car il reste le plus utilisé. La figure n° 3 est un exemple de débit glottique résultat de filtrage inverse et du fitting du modèle LF correspondant. Sur le résultat stylisé par le modèle LF trois paramètres caractérisant le débit glottique sont alors extraits automatiquement (figure n° 4) : les quotients d'ouverture et de dissymétrie Qo et Qd caractérisant la forme et la fréquence fondamentale Fo. Un quatrième paramètre, l'énergie E du signal aux lèvres mesurée sur le signal de parole, vient compléter la caractérisation. Le modèle à deux masses utilisé ayant un retour de phase



$$Q_O = \frac{t_o - t_e}{t_o} \quad ; \quad Q_d = \frac{t_d - t_e}{t_o - t_d} \quad ; \quad F_o = \frac{1}{t_o}$$

Figure n°4 : Caractérisation du débit glottique. Mesures de Qo, Qd et Fo faites sur sa dérivée.

lors de la fermeture de glotte constant, aucun paramètre ne caractérise ce retour de phase (similaire au paramètre t_a du modèle LF).

L'espace de contrôle du modèle des cordes vocales est ainsi constitué des quatres paramètres Qo, Qd, Fo, E. L'espace de commande est réduit aux trois paramètres servant à commander le modèle à deux masses : Ag0, aire d'ouverture de la glotte au repos, Q, facteur masse-tension et Ps pression subglottique.

MODELE DIRECT

L'espace de commande est échantillonné afin de réaliser un dictionnaire de correspondances entre les entrées et les sorties du contrôleur. L'échantillonnage des trois paramètres de commande, choisi pour parcourir toute la plage de fonctionnement du modèle à deux masses dans des conditions de production de voix non pathologiques, est le suivant :

Ag0 : 0.05 ; 0.1 ; 0.15 ; 0.2 ; 0.3 ; 0.4 ; 0.5 ; 0.6 ; 0.7 ; 0.8 cm2

Q : 0.5 ; 1 ; 1.5 ; 2 ; 2.5 ; 3 (sans unité)

Ps : 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12 ; 13 ; 15 cmH_2O.

Pour ces six cent points de l'espace de commande sont calculés les points de l'espace de contrôle correspondants.

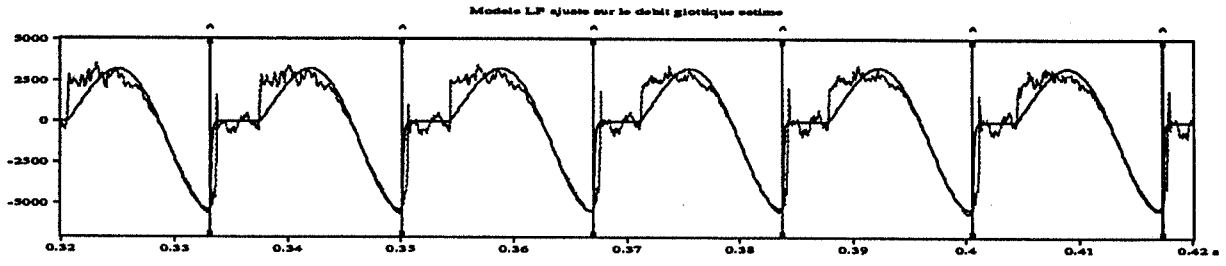


Figure n°3. Résultat de filtrage inverse d'un signal synthétique et fitting du modèle LF

Afin de construire le modèle direct du modèle des cordes vocales une approximation par une fonction polynomiale d'ordre deux est choisie. Cette fonction permet, à partir d'un point de l'espace de commande du modèle, de prédire le point de l'espace de contrôle atteint. Pour améliorer les résultats de l'approximation les paramètres de commandes sont normalisés. L'erreur RMS due à l'approximation est calculée en pourcentage de la variation du paramètre concerné. Elle permet d'évaluer les performances de l'approximation polynomiale pour chaque paramètre de l'espace de contrôle. Cette erreur calculée pour Fo, E, Qo, Qd étant de respectivement 1 %, 3 %, 4 % et 13 %, les paramètres qui se prètent le mieux à cette approximation sont Fo, E et Qo. Le paramètre Qd, à cause de sa plus grande sensibilité aux erreurs de mesure, est par contre mal approximé par son polynôme, il faut donc s'attendre à ce que son évaluation par la méthode d'inversion proposé soit peu précise.

Le contrôleur gestuel est conçu autour d'un algorithme de rétropropagation du gradient (Rumelhart, Hinton & Williams 1986) qui permet de rétropropager l'erreur commise sur les paramètres de l'espace de contrôle vers les paramètres de l'espace de commande. Une séquence d'apprentissage composée de l'association de valeurs d'entrée et de sortie est fournie au contrôleur. L'algorithme se charge alors de minimiser l'erreur commise sur les sorties mais aussi les erreurs portant sur des critères de lissage (figure n°5).

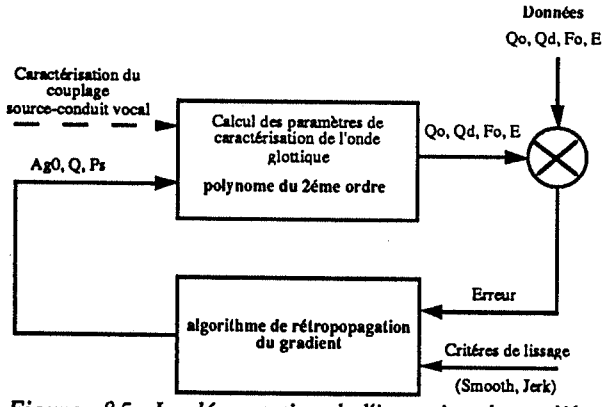


Figure n° 5 : Implémentation de l'inversion du modèle des cordes vocales

VALIDATION

L'utilisation de signaux de synthèse générés par notre modèle de production avec un jeu de paramètres de commande parfaitement connu, s'avère l'option la plus simple pour valider le modèle direct et l'algorithme de rétropropagation du gradient du contrôleur gestuel. De plus, les caractéristiques du débit glottique seront mesurées sur la dérivée de ce dernier directement produite par S.I.M.O.N.D. : les éventuelles erreurs de mesure dues à l'imperfection des techniques de filtrage inverse sont ainsi annulées.

Un premier signal, appelé PRESSION, est synthetisé avec les paramètres suivants : Ag0 et Q sont constants pendant toute la simulation et prennent les valeurs respectivement de 0.05 cm^2 et 3 (sans unité) ; Ps varie linéairement entre 7 et 15 cmH_2O. Un deuxième signal (TENSION) est quant à lui produit avec Ps et Ag0 constants (7.5 cmH_2O et 0.05 cm^2 respectivement) et avec une variation temporelle linéaire du facteur masse tension Q (entre 1 et 2). Les deux signaux ont une durée de 400 ms (divisée en 40 trames de 10 ms chacune). Les quatre paramètres de caractérisation du débit sont lissés avant d'être utilisés dans l'algorithme de rétropropagation du gradient. Dans ces exemples l'erreur est minimisée sur les seuls paramètres Fo, E et Qo ; Qd n'est pas utilisé étant mal approximé par son polynôme. La géométrie du conduit vocal est neutre, correspondant à une configuration vocalique proche du /œ/.

- Inversion du signal PRESSION

Une augmentation de la pression subglottique a pour principal effet d'augmenter l'énergie du signal au lèvres et la fréquence d'oscillation des cordes vocales. Ces deux paramètres sont les moins sensibles aux erreurs de mesure et les plus adaptés à l'approximation polynomiale. La figure n° 6, qui donne l'état du réseau après convergence, montre que les commandes estimées du modèle sont très proches de celles effectivement utilisées pour la simulation. Les valeurs du paramètre Qd, même si elles ne sont pas utilisées pour évaluer les commandes, sont proches de celles mesurées sur le signal de synthèse. Ce paramètre n'est donc pas utile pour inférer les commandes du modèle. Il apparait que dans ces conditions d'utilisation du modèle à deux masses pour cet exemple, le paramètre Qd est fortement corrélé aux autres paramètres de caractérisation du débit glottique.

- Inversion du signal TENSION

Le paramètre important de cette simulation est la fréquence d'oscillation des cordes vocales ; les autres paramètres sont peu affectés par une augmentation de Q, du moins dans cette plage de variation restreinte. Le résultat de l'inversion est donné figure n° 7. On observe notamment une variation de Ps autour de sa valeur cible, mais ces différences sont toujours inférieures à 1 %. L'évolution de la valeur estimée de Ag0 reste totalement monotone bien que légérement supérieure à la valeur effective entrée dans le synthétiseur. Les différences s'expliquent par les erreurs de l'approximation polynomiale. Notre contrôleur gestuel reproduit cependant trés fidèlement l'évolution linéaire du facteur masses-tension Q. Quoique l'évolution des valeurs cibles de Qd soit assez tourmentée (imprécision de la mesure), les résultats de l'inversion montrent des valeurs de Qd proches d'une moyenne de cette évolution : cela confirme la corrélation de ce paramètre aux autres, du moins pour le modèle à deux masses tel qu'il est utilisé ici.

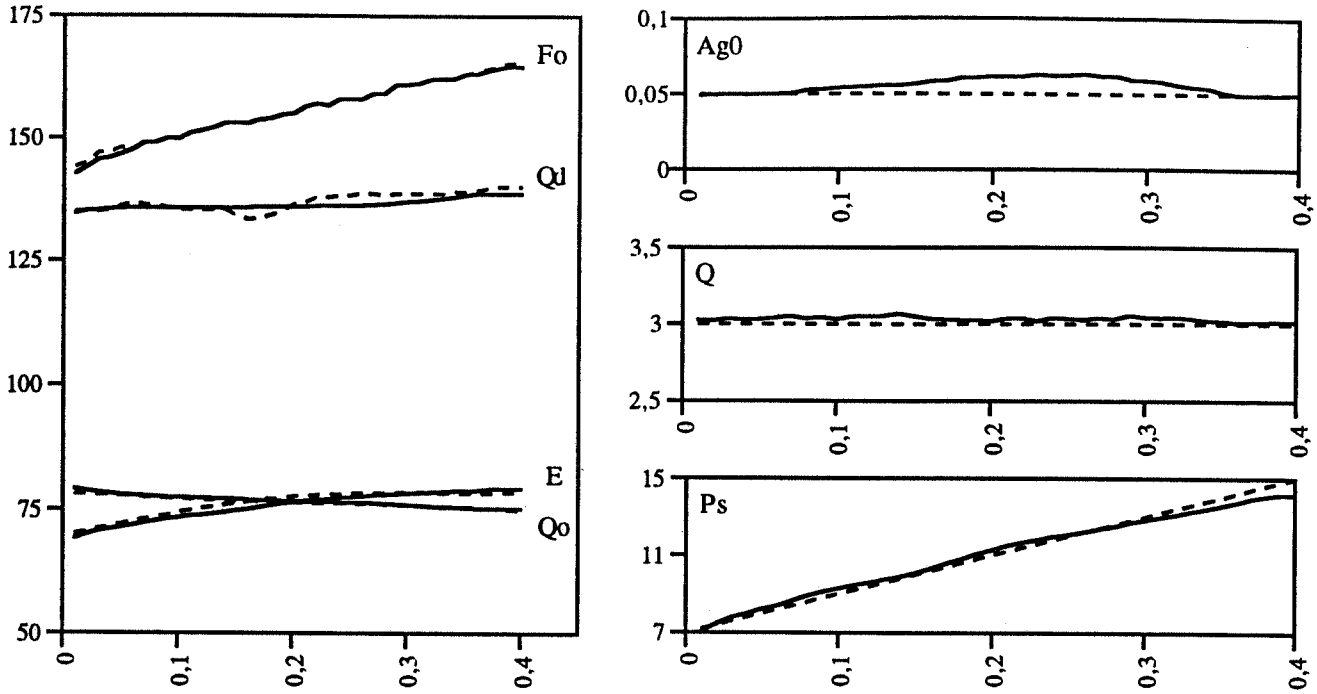


Figure N° 6 : Inversion de PRESSION

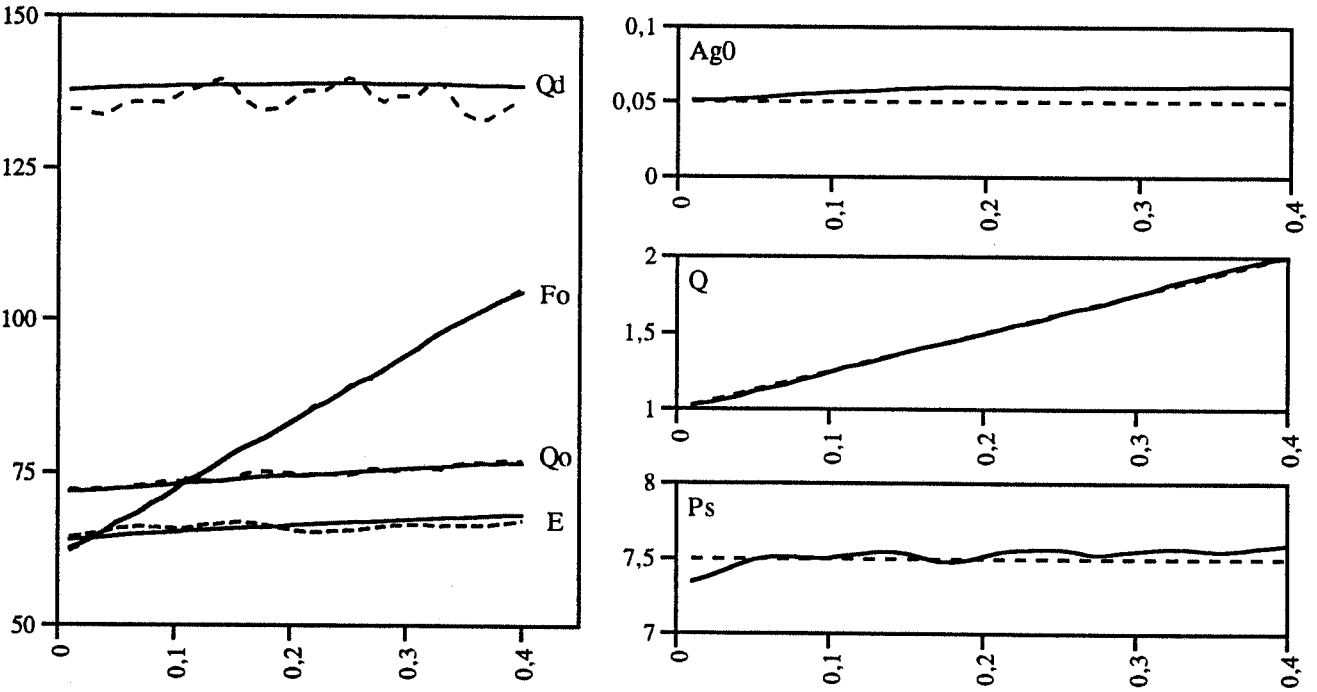
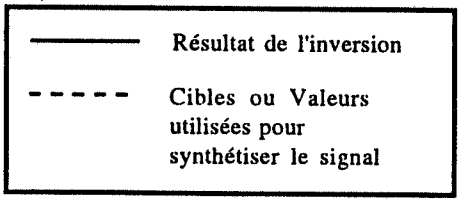


Figure N° 7 : Inversion de TENSION
Pour tous les graphiques: abscisse temporelle en secondes, AgO en cm2, Q sans unité, Ps en cmH2O, Fo en Hz, E en dB, Qo et Qd en %.



INVERSION D'UN SIGNAL DE PAROLE NATURELLE

Un logatome /*pœ*/ prononcé avec trois intensités différentes (faible, moyenne et forte) mais pour une fréquence fondamentale constante, constitue un petit corpus sur lequel notre technique d'inversion des cordes vocales est appliquée. Les signaux de parole naturelle sont enregistrés dans une chambre pseudo-anéchoïque grâce à un micro d'instrumentation Brüel & Kjær n'introduisant aucune distorsion de phase à basses fréquences. Simultanément, une mesure de la pression intra-orale, réalisée grâce à un petit tuyau placé dans le conduit oral et relié à un capteur de pression, permet d'évaluer pendant la production de la plosive /*p*/ la pression subglottique utilisée par le locuteur : pendant le /*p*/ les cordes vocales s'ouvrent et la pression intra-orale est peu différente de la pression subglottique. Sur le signal de parole est évaluée l'énergie E puis, avec la technique de filtrage inverse décrite plus haut, le débit glottique est estimé et caractérisé par fitting du modèle LF pour obtenir Qo, Qd et Fo. La mesure de pression intra-orale fournit une évaluation du paramètre Ps, commande du modèle des cordes vocales.

Avec ces paramètres, l'inversion des cordes vocales est réalisée par notre contrôleur gestuel. Les évolutions des commandes Ps, Q et Ag0 estimées, ainsi que les estimations des paramètres de contrôle comparées aux cibles, sont données ci-dessous (table N°1).

Comme dans les exemples précedents, l'erreur est minimisée sur les seuls paramètres Fo, E et Qo. Les cibles sont très bien atteintes par notre méthode, une légère erreur peut être observée sur les résultats correspondants à l'intensité forte. Ceci s'explique par une mauvaise approximation polynomiale du modèle direct due au fonctionnement non linéaire du modèle à deux masses pour des pressions subglottiques importantes, ainsi que par la qualité médiocre de l'estimation du débit glottique obtenu par filtrage inverse (présence de "ripple"). On remarque que la valeur du paramètre Qd mesuré sur le signal de parole est ici très mal approximée par la valeur donnée par notre méthode. Ce phénomène s'explique par le fait que le modèle à deux masses est un modèle générique des cordes vocale, en ce sens il n'est pas capable "d'imiter" un locuteur donné. Le timbre de voix d'un locuteur étant surtout donné par la forme du débit glottique (donc les paramètres Qo et surtout Qd), il est très difficile pour le modèle à deux masses de reproduire un signal sur des contraintes portant à la fois sur les paramètres Qo et Qd.

La mesure de la pression intra-orale permet de vérifier la validité du paramètre Ps estimé par notre méthode d'inversion. Les estimations sont en effet très proches des valeurs mesurées lors de la phonation, la pression subglottique augmente avec l'intensité du signal de parole. On remarque, de même que la valeur du facteur masse-tension estimé (Q) décroit lorsque l'intensité du signal de parole augmente : le locuteur, en maintenant l'intonation des logatomes constante (fréquence Fo constante) a compensé l'augmentation de Ps en agissant sur la tension de ses cordes vocales. Cette stratégie montrée ici semble raisonnable. Quant au paramètre Ag0 estimé il reste relativement constant autour d'une valeur de 0,05 cm^2 , valeur admise dans la littérature comme étant l'ouverture au repos des cordes vocales pendant la production des voyelles.

CONCLUSIONS ET PERSPECTIVES

Cette étude préliminaire montre que l'approche robotique utilisée pour l'inversion des cordes vocales semble cohérente.

L'inversion des deux signaux de synthèse a permis de vérifier la validité de l'approximation polynomiale du modèle direct et de l'algorithme de rétropropagation du gradient utilisé dans le contrôleur gestuel. Les

	Valeur mesurée sur le signal de parole					Valeur estimée par l'inversion						
Niveaux	Fo	E	Qo	Qd	Ps	Fo	E	Qo	Qd	Ps	Q	Ag0
Faible	127	66	82	215	6,2	127	66	82	132	5,65	2,77	0,062
moyen	125	72	78	270	8,5	125	72	78	136	8,9	2,45	0,074
Fort	130	76	70	235	11,7	130	75,6	73	143	12,3	2,25	0,04

Table N° 1 : Résultat de l'inversion de parole naturelle
(Fo en Hz, E en dB, Qo et Qd en %, Ps en cmH_2O, Q sans unité, Ag0 en cm^2)

évolutions linéaires des commandes sont fidèlement retrouvées avec des erreurs de prédiction faibles. Cependant, le paramètre Qd, à cause des imprécisions de sa mesure, n'est pas bien approximé.

L'inversion du corpus de parole naturelle a montré que notre technique d'inversion pouvait s'appliquer sur des signaux de parole naturelle et que notre contrôleur inférait des stratégies de commande des cordes vocales cohérentes.

Cette étude préliminaire ne porte que sur la production de voyelles. Dans le but d'utiliser notre technique d'inversion pour des consonnes voisées, il sera nécessaire de prendre en compte les phénomènes de couplage entre la source et le conduit vocal. Les valeurs de la fréquence et de la bande passante du premier formant, les meilleurs candidats pour caractériser la charge acoustique du conduit vocal (Trinh V. et al. 91), seront ajoutées dans la construction de notre modèle direct (en pointillé sur la figure n° 5). La charge acoustique du conduit vocal affecte la forme de l'onde glottique ; le quotien de dissymétrie Qd, peu utilisé jusqu'à lors, prendra une importance plus grande dans la caractérisation du débit glottique. Cependant, le choix des paramètres de contrôle Qo et Qd ne semble pas un choix idéal, ces paramètres étant plus representatifs de variations inter-locuteur que de variations intra-locuteurs de la source vocale. Nous envisageons ainsi d'affiner la caractérisation du débit par l'introduction de l'estimation de la pente spectrale de la source.

REMERCIEMENTS

Ces travaux ont été en partie financés par le projet européen ESPRIT/BR n° 6975 Speech Maps.

REFERENCES :

ALKU P., VILKMAN E., LAINE U.K. (1991) *Analysis of glottal waveform in different phonation types using the new IAIF-method.* Proceed. of the XIIth International Congress of Phonetic Sciences, Aix en Provence, 362-365.

BAILLY G., LABOISSIERE R. & SCHWARTZ J.L. (1991) *Formant trajectories as audible gestures : an alternative for speech synthesis.* Journal of Phonetics, 19-1, 9-23.

CASTELLI E. & BADIN P. (1993) *Time and Frequency Domain Acoustic Models of the Vocal Tract.* Speech Maps - Mapping of Action and Perception in Speech - WP2 Deliverable n° 5, Appendix B, 1-25.

FANT G., LILJENCRANTS J., QI-QUANG L. (1985) *A Four-parameter Model of Glottal Flow.* STL-QPSR 4/1985, 1-13.

ISHIZAKA K. & FLANAGAN J.L. (1972) *Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords.* B.S.T.J., 51, 1233-1268.

KELLY J.L. & LOCHBAUM C.C. (1962) *Speech Synthesis* Proc. Stokholm-Speech Communications Seminar - R.I.T. 127-130. and 4th Int. Congr. Acoust., G42.

MAEDA S. (1990) *Compensatory articulation during speech : evidence from the analysis and synthesis of vocal tract shapes using an articulatory model.* Speech Production and Speech Modelling, W.J. Hardcastle & A. Marchal eds., 131-149.

RUMELHART D.E., HINTON G.E. & WILLIAMS R.J. (1986) *Learning Internal Representation by Error Propagation.* in Parallel Distribued Processing : Exploration in the Microstructure of cognition. by RUMELHART D.E. & McCLELLAND J.L. (eds.), 318-362.
Cambridge, MA : MIT Press.

TRINH VAN L., GUERIN B. & CASTELLI E. (1991) *Source-Tract Coupling and the Subglottal System in an Articulatory Synthesizer.* Eurospeech 91, Genova, vol. 1, 267-270.

FORMULATION VARIATIONNELLE DU LIEN ACOUSTICO-ARTICULATOIRE.

Paul JOSPA

INSTITUT DES LANGUES VIVANTES ET DE PHONETIQUE, CP 110,
UNIVERSITE LIBRE DE BRUXELLES,
50 av. F. D. Roosevelt, 1050 Bruxelles, Belgique.

Abstract
A variational formulation of the resonance modes in the vocal tract is presented. This formulation draws a direct and explicit (non linear) link between the first formant frequencies and the parameters of a given area function model. The adopted condition at the labial end fits the open as well as the closed tract at the lips. The variational method gives us analytical expressions for the sensitivity functions and accounts for wall admittance and lip radiation. It allows to make use of direct methods of the variational calculus, which are efficient, time saving, and simple to implement. These advantages are exploited in our works on acoustic-articulatory inverse mapping.

Mots clés: *conduit vocal, condition aux lèvres, modes de résonance, calcul variationnel, modèle acoustique, fonctions de sensibilité.*

1. INTRODUCTION

Les lois physiques gouvernant les premiers modes de résonance dans le conduit vocal peuvent recevoir une formulation variationnelle [Jospa 1972, 1975, 1982/83], qui présente d'importants avantages vis-à-vis de la formulation différentielle classique. La formulation variationnelle facilite l'étude théorique du problème; elle permet l'usage de méthodes directes du calcul des variations —telle la méthode de Rayleigh-Ritz [Gould 1957]—, conduit à des méthodes de calcul efficientes, simples à mettre en oeuvre et économiques en temps de calcul [Jospa 1975]. Un intérêt majeur de cette méthode réside dans le fait qu'un lien algébrique (non linéaire) direct est établi entre les fréquences formantiques et les paramètres d'un quelconque modèle de la fonction d'aire; considéré de manière inverse, ce lien exprime des contraintes acoustiques auxquelles les paramètres du modèle de la fonction d'aire doivent satisfaire pour produire des fréquences formantiques désirées [Jospa 1972, 1974, 1991]. Outre les fréquences formantiques (fréquences des modes de résonance), la formulation variationnelle permet d'obtenir des expressions explicites des fonctions de sensibilité (sensibilité des modes aux variations locales de la fonction d'aire du conduit [Fant et Pauli 1974]) calculables dans des temps très brefs. Ces avantages sont exploités dans le développement de procédures rapides réalisant des transformations acoustico-articulatoires inverses [Jospa, Soquet, Saerens 1992], [George, Jospa, Soquet, 1994]. Dans ce papier, avec la méthode variationnelle, nous présentons une modélisation de l'effet du rayonnement aux lèvres capable de traiter, dans le cadre de la formulation variationnelle, n'importe quel degré d'ouverture aux lèvres, y compris le cas des occlusives labiales. Une extension de la formulation variationnelle applicable au cas des voyelles et consonnes nasales est présentée dans [Jospa 1994 (soumis à publication)].

2. LE MODELE ACOUSTIQUE

2.1. Modèle de Sondhi et mode stationnaire du champ sonore.

Formellement, l'équivalence entre les formulations variationnelle et différentielle —cette dernière s'exprimant par les équations de continuité et du mouvement— n'est strictement établie que dans le cas où le modèle acoustique admet une admittance des parois du conduit vocal distribuée proportionnellement à la fonction d'aire. C'est précisément cette hypothèse qui caractérise le modèle proposé par Sondhi [Sondhi 1974] (voir aussi [Fant 1972]), modèle qui rend compte des effets provoqués par les pertes sur la propagation de l'onde sonore et sur les paramètres formantiques. Les autres hypothèses adoptées, telle celle d'une propagation sonore en front d'onde quasi-plan, sont tout à fait classiques, excepté pour la condition aux lèvres qui sera présentée plus loin.

Il est commode d'exprimer le champ sonore par son potentiel des vitesses $\phi(x,t)$ défini par[1]: $-\partial_x\phi = v(x,t)$, où x est la distance à la glotte , t le temps, et v la vitesse volumique du champ sonore. La pression sonore $p(x,t)$ peut également s'exprimer en terme du potentiel des vitesses: $p(x,t) = \rho\partial_t\phi$, en vertu de l'équation classique du mouvement: $\partial_x p = -\rho\partial_t v$, où ρ désigne la densité de l'air au repos dans le conduit.

Pour un mode de résonance de fréquence $f = \omega/2\pi$, le potentiel des vitesses prend la forme suivante:

$$\phi(x,t) = \psi(x)e^{-\sigma t}\cos(\omega t), \qquad (1)$$

où ψ représente la distribution spatiale d'amplitude de vibration du mode, et σ l'amortissement induit par les pertes dans le conduit vocal. L'amplitude $\psi(x)$ satisfait l'équation *auto-adjointe* suivante:

$$\frac{\partial}{\partial x}\left(A\frac{\partial\psi}{\partial x}\right) + \frac{\omega^2-\omega_p^2}{c^2}A\psi = 0, \qquad (2)$$

laquelle est une équation de Webster exprimée dans le domaine fréquentiel. Dans cette équation, $A(x)$ représente la fonction d'aire, c la vitesse du son dans le conduit vocal (c = 34000 cm/s), et ω_p une correction de fréquence qui exprime l'effet de l'admittance des parois sur les fréquences des modes de résonance ($\omega_p \approx 175 \,.\, 2\pi\ s^{-1}$).

2.2. Conditions aux extrémités du conduit.

A la glotte, l'impédance est supposée infinie, ce qui entraîne une vitesse volumique nulle en $x = 0$:

$$\partial_x\psi(0) = 0 \qquad (3)$$

Désignons par $R_L \approx \sqrt{A(L)/\pi}$ le "rayon" de l'ouverture des lèvres. L'impédance de rayonnement aux lèvres est souvent approchée, dans la bande de fréquence recouvrant les trois premiers formants, par l'expression suivante [Rayleigh 1898, Flanagan 1972]:

$$z_r \equiv \frac{\tilde{p}(L,\omega)}{\tilde{v}(L,\omega)} \approx \left\{\frac{1}{2}\left(\frac{\omega R_L}{c}\right)^2 + j\frac{8}{3\pi}\frac{\omega R_L}{c}\right\}\rho c,$$

valable pour $\omega R_L/c << 1$, où $\tilde{p}(x,\omega)$ et $\tilde{v}(x,\omega)$ désignent les transformées de Fourier respectivement de la pression et de la vitesse. Le terme résistif, en $(\omega R_L/c)^2$, est responsable d'une dissipation d'énergie acoustique; cette dissipation d'énergie n'influence pas la propagation de l'onde dans le conduit: ni les fréquences de résonance, ni les distributions spatiales des modes ne s'en trouvent affectées de manière sensible. Nous pouvons donc négliger ce terme résistif pour le calcul des modes. Le terme réactif en $j\omega R_L/c$, par contre, a pour effet de diminuer les fréquences propres du conduit "non rayonnant" d'un facteur de l'ordre de $3\pi L/(3\pi L + 8R_L)$ [Flanagan 1972]. En fait, ce facteur, tout en restant proche de l'unité, dépend du mode et de la géométrie du conduit. La condition à considérer aux lèvres pour la résolution des équations de propagation et le calcul des modes, se réduit donc à : $(8R_L/3\pi)p_{,x}(L,t) + p(L,t) = 0$, en passant au domaine temporel. En terme du potentiel des vitesses: $\psi(x)e^{-\sigma t}\cos(\omega t)$, nous avons, avec: $R_L \approx \sqrt{A(L)/\pi}$:

$$A(L)\partial_x\psi(L) + q\sqrt{A(L)}\psi(L) = 0, \qquad (4)$$

avec: $q = (3/8)\pi^{3/2}$.

La tête se comporte plutôt comme un baffle sphérique. L'impédance de rayonnement est dans ce cas fonction du rapport du rayon d'ouverture aux lèvres au rayon du baffle. Pour obtenir une condition aux lèvres mieux appropriée au conduit vocal, nous remplaçons dans (4) le facteur q constant par une fonction du rapport de l'ouverture aux lèvres au diamètre moyen D de la tête, soit: $q = q\left(\sqrt{A(L)}/D\right)$. Cette fonction peut être déterminée expérimentalement en optimisant, pour tout l'espace vocalique, l'accord entre les fréquences de résonance calculées et les fréquences formantiques observées en parole naturelle ou obtenues à l'aide d'autres modèles bien établis du rayonnement aux lèvres . Comme première estimation de cette fonction q —le diamètre de la tête étant constant—, nous proposons l'expression suivante:

$$q = a\sqrt{A(L)} + b, \qquad (5)$$

avec: $a = -3.5\text{cm}^{-1}$, et: $b = 35.0$.

La condition à considérer aux lèvres sera donc la condition (4) avec q donné par (5).

Contrairement au traitement classique qui consiste à *imposer* la condition $\psi(L') = 0$ avec

[1] Pour exprimer les dérivées, nous adoptons indifféremment les notations suivantes: $\frac{\partial f}{\partial x}$ ou $\partial_x f$.

$L' = L + \Delta L$ et $\Delta L \approx (3/8)\sqrt{\pi A(L)}$, la condition (4) permet de traiter aussi bien le conduit ouvert que le conduit fermé aux lèvres. En effet, comme nous le verrons plus loin, la condition (4) ne contredit pas la condition appartenant au conduit fermé aux lèvres, soit: $\partial_x \psi(L) = 0$, et permet de passer de manière continue des voyelles ouvertes aux occlusives labiales.

3. METHODE VARIATIONNELLE

3.1. Formulation variationnelle.

Posons:

$$\lambda = \left(\omega^2 - \omega_p^2\right)/c^2, \qquad (6)$$

et considérons l'expression suivante:

$$I[\varphi;\omega] = \int_0^L A\left((\partial_x \varphi)^2 - \lambda \varphi^2\right) dx + q\sqrt{A(L)}\varphi^2(L) \qquad (7)$$

où $\varphi(x)$ est une fonction suffisamment régulière quelconque[2]. Cette intégrale est une fonctionnelle de φ. Le calcul des variations montre que cette fonctionnelle est stationnaire (ici minimale) vis-à-vis de petites variations arbitraires de φ, si et seulement si φ obéit à l'équation de Webster et aux conditions (3) et (4) aux extrémités du conduit. Il n'est pas nécessaire d'imposer aux fonctions φ de satisfaire *à priori* aux conditions (3) et (4) à la glotte et aux lèvres; ces conditions sont automatiquement vérifiées dès lors que φ minimise la fonctionnelle (7). Pour cette raison, les conditions (3)-(4) sont appelées conditions *naturelles* aux limites pour le problème variationnel: $\delta I[\varphi,\omega]/\delta\varphi = 0$. Les solutions de ce problème s'identifient donc avec les distributions spatiales d'amplitude des modes du conduit vocal, soit: $\varphi = \psi(x)$. Notons qu'en raison de l'homogénéité du problème, celui-ci n'admet de solutions que pour des valeurs discrètes de ω égales aux fréquences des modes.

Pour $\varphi \equiv \psi(x)$, le premier terme du membre de droite de (7) représente "l'intégrale d'action" du champ sonore associé à un mode; le second terme représente un terme de charge introduit pour satisfaire la condition (4) aux lèvres. Quand les lèvres sont closes, la fonctionnelle (7) ne doit pas être modifiée. En effet, quand $A(L) = 0$, la fonctionnelle (7) se réduit précisément à: $I_0[\psi,\omega] = \int_0^L A\left((\partial_x \psi)^2 - \lambda\psi^2\right) dx$. Pour cette fonctionnelle I_0, les conditions naturelles aux limites sont précisément: $\partial_x \psi(0) = 0$ et $\partial_x \psi(L) = 0$; ces conditions appartiennent au conduit fermé aux deux extrémités. La formulation variationnelle, et les méthodes de calcul qui en découlent, sont, par conséquent, applicables autant aux occlusives labiales qu'aux voyelles orales.

L'amplitude $\psi(x)$ vérifiant (2), (3) et (4), il est aisé de montrer que:

$$I[\psi;\omega] = 0. \qquad (8)$$

3.2. Calcul des modes de résonance.

Le calcul des modes peut s'exprimer en ces termes:

a) Déterminer les N premières fréquences propres $\omega = \omega_n$ qui autorisent l'existence des N premières fonctions propres $\psi = \psi_n$, $(n = 1,\ldots,N)$ annulant la variation première de la fonctionnelle I définie par (7).

b) Identifier ces fonctions propres parmi un ensemble complet de fonctions $\varphi(x)$ (fonctions d'essais).

En pratique, le choix des fonctions d'essais est limité à un sous-ensemble non complet de fonctions admissibles, de sorte que les résultats obtenus ne sont généralement que des approximations des solutions exactes.

Le problème ainsi formulé peut être résolu par la méthode de Rayleigh-Ritz [Courant et Hilbert 1953, Gould 1966]. Cette méthode fournit un algorithme de calcul particulièrement efficient [Jospa 1975]. La méthode de Rayleigh-Ritz implique le choix d'un ensemble de M fonctions linéairement indépendantes (fonctions coordonnées), soit: $\eta_m(x)$, $(m = 1,\ldots,M)$. Les fonctions d'essais sont alors définies comme combinaisons linéaires de ces fonctions coordonnées, soit: $\varphi(x;c_1 \ldots c_M) = \sum_{m=1}^{M} c_m \eta_m(x)$. La fonctionnelle (7) est alors approchée par une fonction quadratique ordinaire en les paramètres c_m. Le problème variationnel est ainsi ramené à un problème classique de minimisation d'une fonction quadratique. Comme fonctions coordonnées, nous avons choisi les premiers polynômes de Chebyshev en raison de leur comportement aux limites; d'autres choix conviennent, telles des fonctions trigonométriques.

La méthode établit un lien algébrique direct entre les valeurs approchées des premières fréquences de

[2] φ doit être de carré sommable et deux fois différentiable (voir par ex.: [Courant et Hilbert 1953]).

résonance du conduit (fréquences formantiques) et la géométrie du conduit via les quantités :

$$V_{l,m} = \int_0^L A\eta_l\eta_m dx \qquad (9a)$$

et:

$$K_{l,m} = \int_0^L A(\partial_x\eta_l)(\partial_x\eta_m)dx, \qquad (9b)$$

quantités qui, pour un choix donné de fonctions coordonnées, ne dépendent que de la fonction d'aire. Ce lien prend la forme d'une équation polynomiale de degré M en λ (et donc aussi en ω^2):

$$\det\left[K_{l,m} - \lambda V_{l,m}\right] = 0 \quad \text{(équation caractéristique)} \qquad (10)$$

dont les premières racines fournissent les valeurs approchées par excès des premières fréquences de résonance du conduit.

Pour le calcul des trois premiers modes, une précision satisfaisante est atteinte avec un nombre M de fonctions coordonnées relativement peu élevé, soit $M \approx 10$. On notera, par ailleurs, que la méthode ne fait pas appel au calcul de la fonction de transfert. Elle permet ainsi une économie significative en temps de calcul.

Une fois calculées les valeurs propres (les ω_n^2), le calcul des fonctions propres se réduit à la résolution d'un système de $M-1$ équations linéaires pour chacun des modes.

4. SENSIBILITE DES FREQUENCES DES MODES

4.1 Fonctions de sensibilité.

Désignons par $[1_A]$ et $[1_x]$ les unités respectivement d'aire et de longueur. La fonction de sensibilité de la fréquence d'un mode (soit: $S_n(x) \equiv [1_A]^{-1}\,\delta f_n/f_n$) exprime le taux de déplacement de la fréquence provoqué par une faible perturbation de la fonction d'aire située en x et représentée par une distribution delta de Dirac: $\delta A(x) = [1_A]\delta(x'-x)$. L'expression analytique de ces fonctions de sensibilité s'obtient aisément à partir du résultat (8):

$$\int_0^L A\left((\partial_x\psi_n)^2 - (2\pi/c)^2\left(f_n^2 - f_p^2\right)\psi_n^2\right)dx + q\sqrt{A(L)}\left(\psi_n(L)\right)^2 = 0, \qquad (8bis)$$

où f_n désigne la fréquence du n^{ème}. mode, et f_p la correction de fréquence qui rend compte de l'effet de l'admittance des parois sur le mode. Un simple calcul de perturbation du premier ordre appliqué à (8bis) fournit l'expression cherchée :

$$S_n(x) = \frac{\left((c/2\pi)^2(\partial_x\psi_n(x))^2 - \left(f_n^2 - f_p^2\right)\psi_n^2(x)\right)[1_x]}{2f_n^2\int_0^L A\psi_n^2 dx}, \qquad (11a)$$

pour $x \neq L$.

Pour $x = L$, nous avons:

$$S_n(L) = \frac{\left((c/2\pi)^2(\partial_x\psi_n(L))^2 - \left(f_n^2 - f_p^2\right)\psi_n^2(L)\right)[1_x]}{4f_n^2\int_0^L A\psi_n^2 dx} + \frac{(c/2\pi)^2\left(\partial_{A(L)}q + (q/2A(L))\right)\sqrt{A(L)}\psi_n^2(L)}{2f_n^2\int_0^L A\psi_n^2 dx}. \qquad (11b)$$

Les fonctions de sensibilité sont exprimées entièrement en terme des grandeurs non perturbées f_n, $\psi_n(x)$ and $A(x)$; elles s'appliquent au conduit ouvert comme au conduit fermé aux lèvres. Le calcul numérique des fonctions de sensibilité est immédiat, une fois calculés les f_n et $\psi_n(x)$ par la méthode esquissée au paragraphe précédent.

Soit: $\hat{\omega}_n \equiv \sqrt{\omega_n^2 - \omega_p^2}$, la fréquence du nième mode pour le conduit à *parois idéalement rigides*. En terme des densités d'énergie cinétique: $e_{Kin,n}(x) = (\rho/4)A(\partial_x\psi_n)^2$ et potentielle: $\hat{e}_{Pot,n}(x) = (\rho/4)(\hat{\omega}_n/c)^2 A\psi_n{}^2$ du conduit à parois idéalement rigides, l'expression (11a) peut être réécrite sous la forme:

$$S_n(x) = \frac{\left(e_{Kin,n}(x) - \hat{e}_{Pot,n}(x)\right)/A(x)}{\left(1 + (\omega_p/\hat{\omega}_n)^2\right)2\hat{E}_{Pot,n}}[1_x] \qquad (11a\text{ bis})$$

pour $x \neq L$, avec: $\hat{E}_{Pot,n} = \int_0^L \hat{e}_{Pot,n}dx$. Cette expression diffère quelque peu de celle donnée dans [Fant & Pauli 1974]:

$$S_n(x) = \frac{\left(e_{Kin,n}(x) - \hat{e}_{Pot,n}(x)\right)/A(x)}{\hat{E}_{Tot,n}}[1_x],$$

avec: $\hat{E}_{Tot,n} = \int_0^L e_{Kin,n}dx + \int_0^L \hat{e}_{Pot,n}dx$. L'expression des fonctions de sensibilité donnée par Fant et Pauli suppose qu'il y ait équipartition entre les énergies cinétique et potentielle totales du champ acoustique; or dès que les parois ne sont pas parfaitement rigides, une

part de l'énergie potentielle acoustique est absorbée par les vibrations mécaniques des parois. L'équipartition de l'énergie acoustique n'est donc pas réalisée, et, en lieu et place, nous vérifions en réalité, en vertu de (8bis), une égalité du type suivant:

$E_{Kin,n} = \left(1-(\omega_p/\omega_n)^2\right) E_{Pot,n}$ – terme de charge aux lèvres.

La principale différence observée entre notre expression des fonctions de sensibilité et celle donnée par Fant et Pauli réside dans le facteur: $\left(1+(\omega_p/\hat{\omega}_n)^2\right)^{-1}$ (voir 11a bis) qui réduit d'autant les amplitudes de nos fonctions de sensibilité par rapport à celles de Fant et Pauli. Cette réduction d'amplitude est surtout sensible pour le premier formant; elle est négligeable aux modes élevés. *A traitement égal de la condition aux lèvres* (laquelle, conjointement à la géométrie du conduit, conditionne les distributions d'amplitude $\psi_n(x)$ des modes), les *formes* des fonctions de sensibilité ne sont pas modifiées.

Des exemples de fonctions de sensibilité obtenues par la méthode variationnelle sont illustrées par les Figures 1a-d.

4.2. Sensibilité aux paramètres d'un modèle articulatoire.

Désignons par $A(x;\alpha_1,...,\alpha_K)$ la fonction d'aire générée par un modèle articulatoire (ou modèle de la fonction d'aire) quelconque, contrôlé par K paramètres articulatoires $\alpha_1,...,\alpha_K$. La sensibilité des fréquences formantiques aux petites variations des paramètres articulatoires, s'exprime à l'aide du Jacobien de la transformation acoustico-articulatoire; le Jacobien est défini par la matrice J de dimensions NxK, dont les éléments sont définis par:

$$J_{n,k} = \frac{\partial f_n}{\partial \alpha_k} \qquad (12)$$

N étant le nombre de fréquences formantiques considéré ($N=3$). Etant donné la définition des fonctions de sensibilité (voir §4.1), les éléments de matrice du Jacobien peuvent s'exprimer sous la forme:

$$J_{n,k} = f_n \int_0^L S_n(x) \frac{\partial A(x)}{\partial \alpha_k} dx,$$

$$n = 1,...,N \quad \text{et} \quad k = 1,...,K \; . \qquad (13)$$

Le Jacobien est ainsi aisément calculable, une fois obtenues les fonctions de sensibilités; le calcul ne fait intervenir que les grandeurs acoustiques f_n et $\psi_n(x)$ non perturbées.

5. CONCLUSION

La formulation variationnelle, telle que nous l'avons présentée, nous a permis de développer une méthode de calcul des modes de résonance —basée sur la méthode de Rayleigh-Ritz— précise et particulièrement rapide. La modélisation de l'effet du rayonnement aux lèvres que nous avons introduite, et son traitement dans le cadre de la formulation variationnelle, a l'avantage de s'appliquer aussi bien au conduit fermé qu'au conduit ouvert aux lèvres. Les expressions explicites des fonctions de sensibilité sont aisées à calculer, et s'appliquent indifféremment au conduit ouvert comme au conduit fermé aux lèvres. Il en est de même pour le calcul des Jacobiens des transformations entre niveaux acoustique et articulatoire, dont l'utilité pour résoudre des problèmes d'inversion acoustico-articulatoire est évidente.

Remerciements

Ce travail a été supporté partiellement par un contrat de la Communauté Européenne: Contrat SCIENCE PROGRAM, No. SC1*-CT92-0786.

BIBLIOGRAPHIE.

COURANT R. & HILBERT D. (1953): Methods of mathematical physics. Interscience Publ. Inc., New York, Vol.1.

FANT G. (1960): Acoustic theory of speech production. Mouton & Co., 's Gravenhage.

FANT G. (1972): "Vocal tract wall effects, losses, and resonance bandwiths", STL-QPSR 2-3, pp 28-52.

FANT G. & PAULI S. (1974): "Spatial characteristics of vocal tract resonance modes", in: Speech Communication., Ed.: G, Fant, Proc. Speech Comm. Seminar, Stockholm, Aug 1-3, 1974, Vol.2, pp 121-132.

FLANAGAN J.L. (1972): Speech Analysis Synthesis and Perception. Springer-Verlag, Berlin-Heidelberg-New York.

GEORGE M., JOSPA P. & SOQUET A. (1994): "Generation de transitions articulatoires à l'aide d'un réseau de neurones". Actes des 20èmes J.E.P., Trégastel.

GOULD S.H. (1966): Variational Methods for Eigenvalue Problems, London: Oxford University Press.

JOSPA P. (1972): "Forme approchée du conduit vocal déduite des fréquences de résonance. Théorie des perturbations et méthode variationnelle", Lannion: JEP, GALF, 225-262.

JOSPA P. (1974): "On the use of critical formant frequencies in designing an articulatory model". in: Speech Communication., Ed.: G, Fant, Proc. Speech Comm. Seminar, Stockholm, Aug 1-3, 1974, Vol.2, p.285-292.

JOSPA P. (1975): "Calcul des premières fréquences de résonance du conduit vocal par la méthode variationnelle". Rapp. d'Activités Inst. de Phon. - ULB, 9/12, pp.59-76.

JOSPA P. (1982/83): "Theory of Evolving Normal Modes and the Vocal Tract. - Part I: Basic Relations and Adiabatic Invariance". Acustica, 52, 2, pp. 86-94

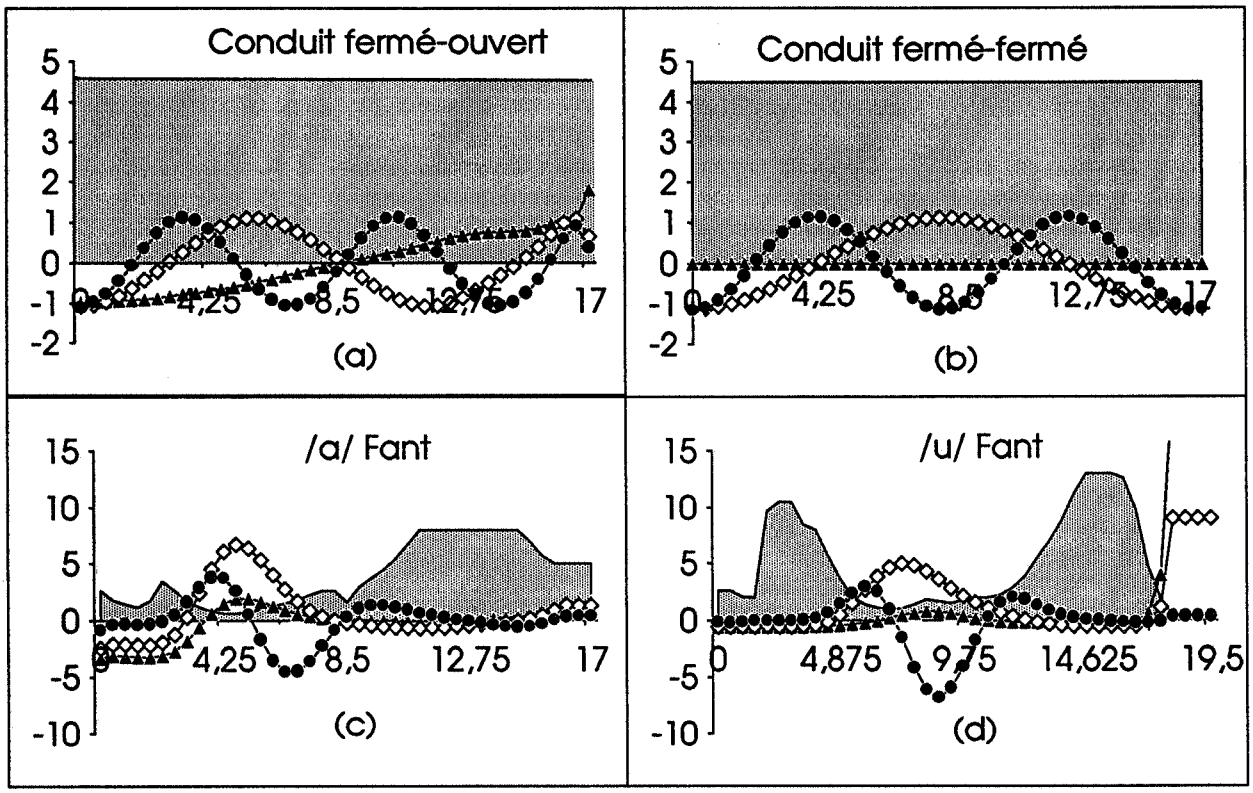
JOSPA P. (1991): "Des paramètres formantiques au profil articulatoire.", Aix-en-Provence: Actes du XIIème Congrès International. des Sciences Phonétiques, vol. 2, pp. 378-381..

P. JOSPA (1994). "Variational characterisation of the resonance modes in the vocal tract". Soumis à publication dans Acta Acustica.

JOSPA P., SOQUET A., & SAERENS M. (1992): "Acoustical sensitivity functions and the control of a vocal tract model", dans Signal Processing VI: Proc. of EUSIPCO-92, Bruxelles, vol 1, pp 171-174.

RAYLEIGH J.W.S. (1898): The theory of sound. (Réédition: Dover Publ., New York, 1945).

SONDHI M.M. (1974): "Model for wave propagation in a lossy vocal tract". J. Acoust. Soc. Am., 55, 5, pp 1070-1075.



Figures 1a-d: *Fonctions de sensibilité des trois premières fréquences formantiques calculées par la méthode variationnelle selon les équations (11a-b).*
En abscisse: la distance à la glotte (en cm).
En ordonnée: la fonction d'aire du conduit (en cm^2), et les fonctions de sensibilité (multipliées par deux pour plus de clareté; en $cm^{-2}hz^{-1}$).
Fonctions de sensibilité : triangles pleins: premier formant $S_1(x)$, losanges: deuxième formant $S_2(x)$, cercles pleins: troisième formant $S_3(x)$. En gris: la fonction d'aire.
a): Conduit uniforme ouvert aux lèvres.
b): Conduit uniforme fermé aux lèvres.
c): voyelle /a/ [Fant 1960].
d): voyelle /u/ [Fant 1960].

ADAPTATION D'UN MODÈLE ARTICULATOIRE À UN LOCUTEUR DANS LE BUT DE CONTRAINDRE L'INVERSION ARTICULATORI-ACOUSTIQUE

Denis BEAUTEMPS et Bernard GABIOUD

LABORATOIRE D'ANALYSE INFORMATIQUE DE LA PAROLE (LAIP),
UNIVERSITÉ DE LAUSANNE — CH-1015 LAUSANNE

(Denis.Beautemps, Bernard.Gabioud)@imm.unil.ch

Résumé

Cette étude présente l'utilisation d'un modèle articulatoire comme contrainte majeure sur les formes articulatoires d'un locuteur masculin (que nous désignerons par PiB dans cet article) dans le cadre de l'obtention de données articulatoires par inversion de la relation articulatori-acoustique.

Le problème de normalisation entre ce sujet et le modèle articulatoire est résolu par l'application au conduit vocal d'une transformation géométrique obtenue par la comparaison des coupes sagittales disponibles à la fois pour le locuteur PiB et pour le modèle.

L'inversion se base sur la minimisation d'une fonction d'erreur sur la fréquence des formants et bénéficie du petit nombre de paramètres du modèle. Le phénomène de compensation entre les paramètres articulatoires nécessite des contraintes supplémentaires.

Mots-clés: *relation articulatori-acoustique, modèle articulatoire, normalisation, coupes sagittales, inversion, minimisation d'erreur, contraintes.*

1. INTRODUCTION

Il est bien connu que, pour effectuer des progrès importants dans le domaine de la production en parole, il est nécessaire d'acquérir d'importantes quantités de données articulatoires. Des mesures directes du conduit vocal à l'aide de techniques d'imagerie telles que les rayons X ou la RMN (Baer et al., 1991), ainsi que des mesures partielles obtenues par électropalatographie ou articulométrie magnétique (Perkell et al., 1992) sont de bons moyens pour y parvenir.

Une autre approche est d'inférer ces données à partir du signal acoustique associé (Schroeder, 1967). Si l'on dispose d'un bon modèle direct de production de la relation articulatori-acoustique, fiable dans ses performances statiques et dynamiques, il est possible de l'inverser pour obtenir des données sur la configuration du conduit vocal à partir du signal acoustique (Flanagan Ishizaka & Shipley, 1980).

Cependant il est connu depuis longtemps que les relations articulatori-acoustiques ne sont pas biunivoques (Atal et al., 1978) et que ce problème d'inversion de la relation articulatori-acoustique tombe dans la catégorie des problèmes mal posés au sens d'Hadamard (Lavrentiev, 1967), c'est-à-dire qu'il n'est pas certain que la solution existe, qu'elle soit unique et qu'elle dépende continûment des conditions initiales.

Dans leur étude sur la récupération de la fonction sagittale et de la fonction d'aire à partir des trajectoires de formants, Beautemps, Badin et Laboissière (1993) ont partiellement résolu le problème de multiplicité de solutions en régularisant le problème par l'utilisation de contraintes de continuité spatiale et temporelle de la fonction sagittale.

Dans le but d'accumuler des données articulatoires cohérentes pour le locuteur PiB, la motivation de ce travail a été (1) de résoudre de manière plus générale le problème de multiplicité des solutions; (2) d'obtenir des fonctions sagittales et fonctions d'aire pouvant être par la suite comparées à des mesures directes du conduit vocal du même sujet.

Tout ceci passe par la mise en oeuvre d'un bon modèle direct valable pour le locuteur PiB pour lequel un certain nombre de données cohérentes articulatoires et acoustiques d'une part et aérodynamiques d'autre part sont déjà disponibles (Badin, 1991).

2. LE MODÈLE DIRECT

Le modèle direct de production acoustique de la parole peut être décomposé en trois modules successifs:

(1) un modèle articulatoire qui fournit une coupe sagittale à partir d'un nombre restreint de paramètres de commande,

(2) un modèle de passage de la fonction sagittale à la fonction d'aire,

(3) un modèle acoustique/aérodynamique incluant la génération de sources d'excitation et la propagation d'ondes planes acoustiques à partir d'une fonction d'aire (Badin & Fant, 1984).

L'utilisation d'un modèle articulatoire (Maeda 1979, 1988) avec un nombre réduit de paramètres est un choix naturel: ayant été obtenu à partir de nombreuses données du conduit vocal d'un sujet humain, il comporte implicitement des contraintes liées à la morphologie de ce conduit. Un nombre réduit de paramètres définit un espace réduit de formes articulatoires. C'est une contrainte majeure dans le cadre de l'inversion du modèle direct.

D'autre part, un modèle articulatoire a l'avantage de gérer de façon cohérente les longueurs et les distances sagittales. Le problème du contrôle des variations de longueur rencontré par Beautemps, Badin et Laboissière (1993) ne se pose donc plus.

Un tel modèle est le modèle articulatoire d'un locuteur bien précis, le sujet PB (Maeda, 1990). Notre intention est de l'adapter au locuteur PiB, pour pouvoir exploiter le domaine articulatoire bien défini du modèle comme contrainte principale de son espace articulatoire.

3. LE MODÈLE ARTICULATOIRE

Pour cette étude, nous avons utilisé une variante du modèle de Maeda de 1988, basé sur l'analyse de cinéradiographies de l'Institut de Phonétique de Strasbourg (Bothorel et al., 1986). Le corpus contient 10 phrases courtes prononcées par le locuteur féminin PB, soit 519 clichés à raison de 50 clichés par seconde.

Le profil sagittal est mesuré à l'aide d'une grille semi-polaire, produisant 25 coordonnées (Figure 1). Suivant le modèle proposé par Lindblom & Sundberg en 1971, ce sont les déviations par rapport à la forme neutre qui sont modélisées.

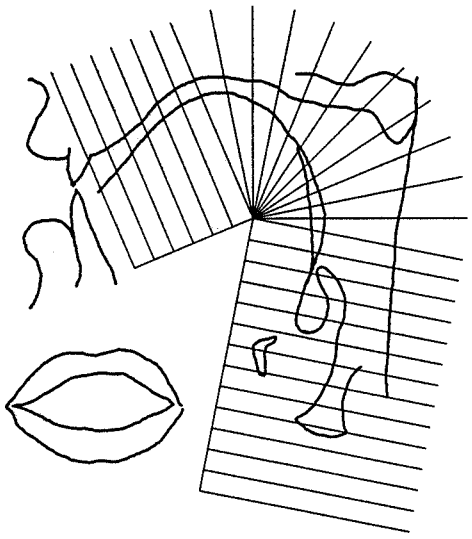


Figure 1. Le tracé manuel d'une cinéradiographie du conduit vocal, avec la grille semi-polaire de Maeda de 1990. En bas à gauche, la vue frontale des lèvres.

Le premier facteur extrait par Maeda est celui de la position de la mâchoire. L'effet de ce paramètre JH est soustrait des données. Ensuite, une analyse en composantes principales des données de la langue permet de dégager 3 facteurs, qui se trouvent être physiquement interprétables: TB, position avant-arrière du corps de la langue, TD, hauteur du dos de la langue et TT, élévation de la pointe de la langue. Ces facteurs correspondent bien aux articulateurs de base de la langue décrits par exemple par Öhman (1966) et Coker (1976).

Les données des lèvres, y compris l'ouverture horizontale, peuvent être adéquatement décrites par 2 paramètres: l'ouverture verticale LH et la protrusion LP. La hauteur du larynx est suivie par le paramètre LY.

Ainsi, le choix de la valeur des 7 paramètres permet de reproduire n'importe quelle forme de conduit vocal compatible avec le corpus de PB.

Bien que résultant de l'analyse d'un conduit vocal féminin, le modèle de Maeda a été adapté à la production de voyelles françaises, telles que prononcées par un locuteur masculin. Pour ce faire, Boë, Perrier & Bailly (1992) ont simplement appliqué un coefficient d'agrandissement global (excepté pour la longueur du palais dur).

En tentant de resynthétiser les phrases originelles utilisées dans l'analyse de Maeda, nous avons découvert une imprécision de la reproduction de la position de l'extrémité de la langue par le modèle: alors qu'au milieu de la langue l'écart-type résiduel est de 0.7 mm, le bout de la langue n'est reproduit qu'avec une précision 3 fois moindre. Une différence de ±2 écarts-types (95% des données) correspond dans la configuration du [y] à une erreur de ±200 Hz sur les formants F_2 et F_3. Cette propriété du modèle factoriel est due à une corrélation locale des données beaucoup plus forte vers le milieu de la langue qu'aux extrémités (Gabioud, 1994).

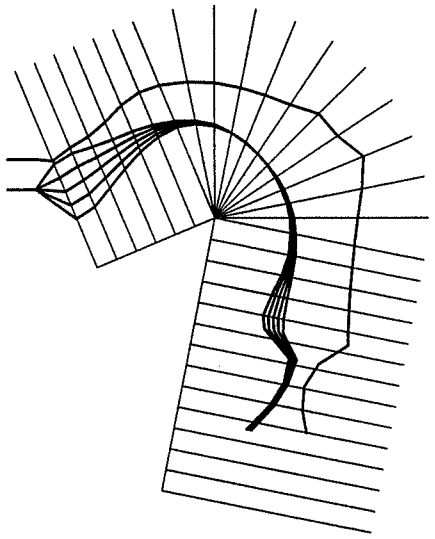


Figure 2. L'effet du nouveau paramètre d'extrémité de la langue, variant de –2 à +2 écarts-types. Il y a un léger effet dans le bas du pharynx, mais beaucoup moins prononcé que dans le modèle de Maeda (1988).

En fait, comme Maeda lui-même le montre (1979), rien n'oblige d'utiliser des composantes principales comme facteurs: on peut choisir des facteurs arbitraires, pour autant qu'ils aient une bonne interprétation physique, de manière à pouvoir leur associer un comportement dynamique. C'est ce que Maeda avait fait dans son modèle de 1979 pour l'ensemble des facteurs de la langue. Suivant ce raisonnement, nous avons construit une variante du modèle de Maeda, avec 3 facteurs pour reproduire la forme de la langue et ayant la même interprétation physique que ceux de Maeda, mais garantissant une meilleure précision à l'avant du conduit vocal. Les deux paramètres définissant la position du corps de la

langue sont déterminés à partir d'une analyse factorielle excluant la pointe de la langue et le larynx. Une fois l'effet de ces paramètres déduits, un troisième facteur décrit la déviation du bout de la langue avec une résolution de ±1 mm (Figure 2).

4. NORMALISATION DU LOCUTEUR

L'adaptation d'un modèle articulatoire à d'autres locuteurs est un problème ancien et difficile (voir par exemple Payan & Perrier, 1993). Il s'agit d'établir une transformation permettant de projeter les espaces acoustique et articulatoire du locuteur étudié dans ceux du modèle, considéré comme le locuteur de référence pour lequel les caractéristiques articulatoires et acoustiques sont bien établies.

Une des approches possibles serait d'inférer des informations sur les longueurs du conduit vocal à partir des formants (Payan & Perrier, 1993). Cependant, du fait que nous disposons de coupes sagittales aussi bien pour le locuteur PiB que pour le modèle, nous sommes en mesure d'établir une transformation géométrique permettant d'adapter les formes articulatoires du conduit de PiB à celles produites par le modèle.

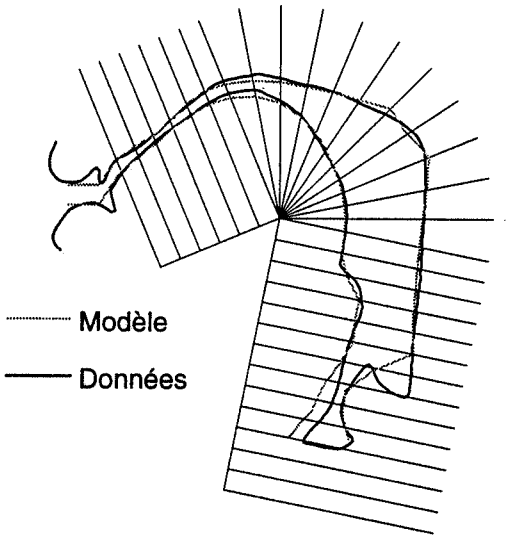


Figure 3. Coupe sagittale transformée du locuteur PiB, avec le résultat de l'ajustement du modèle. Configuration du [i].

Le corpus utilisé contient des coupes sagittales obtenues à partir de clichés radiographiques, et les fonctions sagittales associées pour 3 voyelles [a, i, u] et 6 consonnes fricatives [f, θ, s, ʃ, ç, x] prononcées par le locuteur PiB pour des sons soutenus. Des vues de face des lèvres sont également disponibles (Badin, 1991).

Le fait que ce corpus contienne des données géométriques sur des configurations aussi différentes en termes de lieu et forme de constriction devrait assurer que la normalisation obtenue soit assez générale pour ce locuteur.

En comparant les fonctions sagittales de six configurations disponibles dans les corpus des deux locuteurs, on constate qu'un facteur d'homothétie de 0.80 appliqué sur les données de PiB rend approximativement égales, pour le même phonème, les longueurs totales des conduits. Après l'application de ce facteur, les distances sagittales sont très similaires, et la forme des conduits sont semblables, à l'inclinaison du palais dur près.

Pour chacune des configurations, une transformation des données de PiB sans variation significative de longueur a été effectuée de manière à faire coïncider le contour extérieur du conduit avec celui de PB. La principale modification consiste à effectuer une rotation du palais dur de 11° vers le bas. Cette transformation de la coupe sagittale est conçue de manière à ne pas affecter significativement la fonction sagittale.

Il devient donc possible d'appliquer le modèle de PB aux conduits ainsi transformés de PiB. Pour estimer par le modèle une coupe sagittale de PiB, la démarche est la suivante:

(1) placer la grille sur la coupe sagittale transformée (Figure 3),

(2) à partir des distances mesurées sur la grille, inverser le modèle articulatoire, c'est-à-dire obtenir les paramètres du modèle,

(3) générer une coupe sagittale à partir des paramètres ainsi déterminés,

(4) appliquer la transformation inverse de celle appliquée aux données.

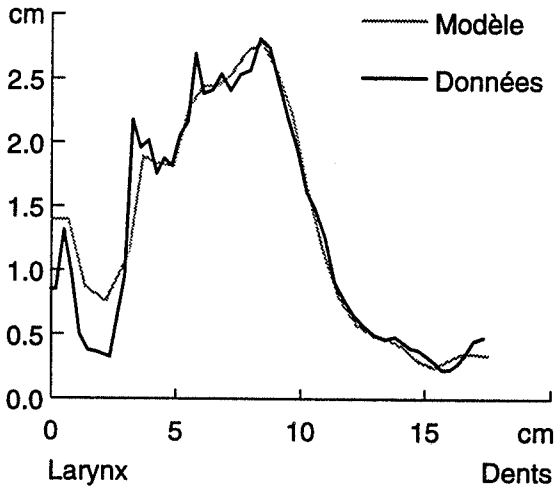


Figure 4. Comparaison de la fonction sagittale mesurée sur le locuteur PiB, dans la configuration du [i], avec la fonction sagittale générée par le modèle.

La figure 4 compare la fonction sagittale de la configuration du [i] avec le résultat de l'ajustement. Les différences sont négligeables, à l'exception de la région du larynx, qui n'a pas été modélisée en détail.

Il faut cependant noter que l'extrémité de la langue n'est pas bien modélisée lorsque celle-ci est en retrait et n'a pas d'intersection avec la dernière ligne de la grille. Mais ce problème existe déjà dans le corpus de PB (Gabioud, 1994) et sort du cadre de ce travail.

5. LA MÉTHODE D'INVERSION

Le problème de normalisation du locuteur PiB est donc en principe résolu. Il est ainsi possible de contrôler les fonctions sagittales du conduit vocal de PiB avec les paramètres, en nombre réduit, du modèle articulatoire.

L'inversion de la relation articulatori-acoustique consiste à déterminer les paramètres du modèle articulatoire à partir du signal acoustique. Ceci est réalisé en minimisant l'erreur calculée sur les formants et pouvant inclure un certain nombre de contraintes supplémentaires.

Le terme d'erreur sur les 5 formants utilisés est donné par la formule:

$$E = \sum \left(\frac{F_i - F_i^*}{S_i} \right)^2$$

où F_i est le formant calculé, F_i^* le formant cible, et S_i est une estimation de l'écart-type de chaque formant mesuré. Dans cette étude, nous avons choisi 10, 20, 40, 80, 120 Hz pour S_1 à S_5 respectivement.

De manière à éviter que les valeurs des paramètres quittent un domaine de définition raisonnable, des limites sur les paramètres ont été introduites sous la forme de termes exponentiels additionnés à l'erreur calculée sur les formants, et pénalisant les solutions interdites. Étant donné que la technique de gradient est utilisée, il est nécessaire que la fonction d'erreur soit dérivable. L'effet de ces termes est négligeable dans le domaine de variation des paramètres du corpus de PB et croît très rapidement à l'extérieur.

De même, pour éviter que la technique de minimisation ne génère des occlusions intempestives lors de l'inversion de voyelles ou de fricatives, des termes exponentiels appliqués aux aires sont introduits dans la fonction d'erreur.

La technique de minimisation utilisée est celle de James & Roos (1975) qui consiste à approcher le minimum par une technique de simplexe puis par une méthode de gradient. La fonction d'erreur est approximée par une forme quadratique dont le minimum est calculé, puis on procède par itération jusqu'à ce qu'un critère de convergence soit satisfait. Cette méthode a aussi l'avantage d'optimiser le nombre d'appels au modèle direct.

Si la solution était unique, cette méthode convergerait à tous les coups. Ce n'est pas notre cas: différentes conditions initiales peuvent donner lieu à des solutions différentes. Les minima locaux ne sont pas forcément aussi optimaux que la solution correcte, mais ils ne sont pas exclus par la méthode de minimisation, d'où la nécessité de critères supplémentaires.

Un aspect intéressant de la méthode de James & Roos est de fournir une matrice de corrélations des erreurs sur les paramètres. Lorsqu'on se trouve au minimum de la fonction d'erreur, un écart de S_i sur la valeur du formant F_i se traduit par une augmentation de 1 de la fonction d'erreur. Plus généralement, si les S_i sont les écarts-types sur les formants, l'équation $E = E_{min} + 1$ définit l'ensemble des solutions situées à 1 écart-type du minimum (Figure 5).

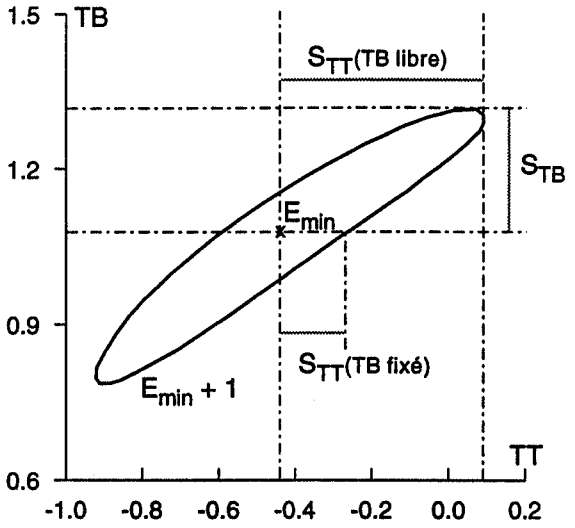


Figure 5. Corrélation entre les erreurs des paramètres TB et TD dans le cas d'un [u]. L'erreur S_{TT} sur le paramètre TT est près de 3 fois plus élevée lorsque TB est variable que lorsque TB est fixé.

6. RÉSULTATS DE L'INVERSION

Afin de tester les performances de cette méthode d'inversion, il est nécessaire d'utiliser une phrase de synthèse générée par le modèle direct, pour pouvoir comparer les paramètres reconstruits par inversion avec les paramètres originels. Dans le cas des phrases réelles, nous n'avons pas cette possibilité.

Notre phrase-test est la deuxième phrase du corpus de PB, soit "Voilà des bougies" [vwaladebuʒi]. Les paramètres originels sont obtenus par inversion géométrique du modèle articulatoire à partir des cinéradiographies.

L'inversion consiste à reconstituer les paramètres du modèle articulatoire à partir des trajectoires des 5 formants associés, échantillonnés toutes les 20 ms. Pour chaque instant, les paramètres obtenus lors de l'inversion à l'instant précédent sont utilisés comme condition initiale de la minimisation. Cette procédure ne constitue pas vraiment une contrainte temporelle mais a l'avantage de produire une cohérence temporelle de la solution obtenue. Il n'y a ainsi aucun risque que le système saute brusquement d'une solution articulatoire à une solution éloignée (par exemple d'un [u] avant à un [u] arrière).

Dans le cas d'un ajustement simultané des 7 paramètres du modèle, on observe deux types de résultat: soit des solutions correctes quant à la forme du conduit mais comportant des erreurs sur les paramètres pouvant atteindre 3 écarts-types, soit des solutions associées à des formes de conduit totalement différentes, bien que les fréquences attendues des formants soient atteintes. Ces

résultats sont d'évidence dus au phénomène bien connu de compensation acoustique entre les articulateurs.

Comme Maeda l'avait observé en 1990, différentes ouvertures de mâchoire sont compensées par un ajustement de la position de la langue et des lèvres. De plus, les deux paramètres de lèvres LH et LP ainsi que la hauteur du larynx LY ont un effet direct sur la longueur du conduit et sont ainsi corrélés. Enfin, la variation d'un paramètre comme LH n'a plus d'effet acoustique au-delà d'une certaine limite.

En conséquence, il est nécessaire de fournir des contraintes supplémentaires. Notre choix a été de fixer certains paramètres à leurs valeurs originelles, ce qui équivaut à utiliser des valeurs mesurées dans le cas de l'inversion d'un signal réel.

Si l'on fixe les paramètres JH, LH et LP aux valeurs originelles, on obtient une solution stable et correcte pour l'ensemble de la phrase test, à l'exception du [u], où l'on observe une instabilité due à une compensation entre les paramètres TB, TD et TT, qui conserve la forme du conduit: une erreur de -0.48 sur TB est compensée par un écart de -0.36 sur TD et de -0.08 sur TT, alors que l'erreur observée sur l'ensemble des paramètres pour le reste de la phrase ne dépasse par 0.01 écart-type.

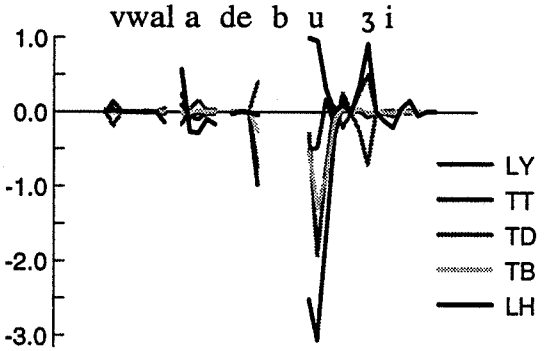


Figure 6. Écart observé sur les paramètres, entre les valeurs obtenues par inversion et les valeurs originelles, avec 5 paramètres variables: LH, TB, TD, TT et LY.

En permettant au paramètre LH d'être ajusté lui aussi, on observe une très forte instabilité de la solution du [u], due aux fortes corrélations entre les erreurs sur les paramètres TT, TD, TB et LH (Figure 6). Un autre cas de compensation est observé dans le [ʒ], où une ouverture des lèvres trop faible (ce qui équivaut à un tube plus long) est compensée par une élévation du larynx et un déplacement de la constriction vers l'avant grâce à un ajustement des paramètres de la langue.

L'inversion de la relation articulatori-acoustique reste un «problème mal posé», en ce sens que ni l'existence de la solution, ni son unicité et sa continuité ne sont démontrées.

Cependant la méthode proposée produit toujours une solution plus ou moins proche de la cible formantique (résolvant à sa manière le problème d'existence), et, dans la mesure où la configuration initiale pour la minimisation de l'erreur est proche de la cible, la méthode garantit qu'on en restera proche au cours du temps (résolvant la continuité tout en ignorant la question de l'unicité).

Cependant, dans la mesure où l'on ne cherche pas une solution exacte mais que l'on se contente d'une certaine approximation, la trajectoire inversée sera définie par une suite de points dans les ellipsoïdes d'erreur (Figure 5).

C'est seulement dans le cas où les corrélations des erreurs sont importantes (forte compensation articulatoire), qu'il sera nécessaire de disposer de contraintes supplémentaires.

7. CONCLUSIONS ET PERSPECTIVES

Dans le cadre de la modélisation articulatoire de la production de la parole, nous disposons du modèle articulatoire de Maeda, qui est contrôlé par un nombre réduit de paramètres, mais modélise un locuteur particulier. Désirant un modèle pour un autre locuteur, nous devons adapter le modèle.

La comparaison entre les formes du conduit vocal fournies par le modèle et celles du locuteur PiB ont permis de déterminer une transformation de la coupe sagittale permettant d'utiliser le modèle de Maeda pour ce locuteur.

Dans le cadre de ce travail, nous avons produit une variante de ce modèle comportant un meilleur paramètre pour décrire le mouvement du bout de la langue.

La multiplicité des solutions caractéristique de l'inversion de la relation articulatori-acoustique est très largement réduite par l'utilisation d'un modèle ne comprenant qu'un nombre restreint de paramètres.

Il est clairement établi que la seule utilisation des fréquences de formants dans la minimisation de l'erreur ne suffit pas à rendre unique et continue la solution d'une inversion simultanée de tous les paramètres du modèle. Cependant, si l'on utilise des valeurs mesurées pour les paramètres de lèvres et de mâchoire, l'inversion fonctionne de manière satisfaisante.

Le phénomène de compensation observé entre les articulateurs nécessite l'introduction de contraintes supplémentaires qui peuvent consister soit en une régularisation temporelle de la solution, soit en l'utilisation d'un modèle dynamique de l'évolution des paramètres (Perrier, Abry & Keller, 1989; Lœvenbruck & Perrier, 1993).

Lorsque que ce travail complémentaire sera effectué, l'utilisation de signaux acoustiques pour générer des données articulatoires sur le sujet PiB à partir du signal acoustique deviendra possible.

RÉFÉRENCES:

Atal, B.S., Chang, J.J., Mathews, M.V., & Tukey, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America, 63,* 1535-1555.

Badin, P. (1991). Fricative consonants: acoustic and X-ray measurements. *Journal of Phonetics, 19,* 397-408.

Badin, P., & Fant, G. (1984). Notes on vocal tract computation. *STL-QPSR 2-3/1984,* 53-108.

Baer, T., Goore, J.C., Gracco, L.C., & Nye P.W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America, 90,* 799-828.

Beautemps, D., Badin, P., & Laboissière, R. (1993). Recovery of vocal tract midsagittal and area functions from speech signal for vowels and fricative consonants. In *Proceedings of the 3rd European conference on Speech communication and Technology, Vol. 1* (pp. 73-76). Berlin.

Boë, L.-J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production : Proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics 20,* 27-38.

Bothorel, A., Simon, P., Wioland, F., & Zerling J.-P. (1986). *Cinéradiographie des voyelles et des consonnes du français. Recueil de documents synchronisés pour quatre sujets: vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques.* Strasbourg, France: Institut de Phonétique.

Coker, C. (1976). A model of articulatory dynamics and control. *Proc. IEEE, 64,* 452-460.

Flanagan, J.L., Ishizaka, K., & Shipley, K.L. (1980). Signal models for low bit-rate coding of speech. *Journal of the Acoustical Society of America, 68,* 780-791.

Gabioud, B. (1994). Articulatory models in speech synthesis. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition* (in press). Chichester: John Wiley.

James, F. & Roos, M. (1975). MINUIT - A system for minimization and analysis of the parameter errors and correlations. *Computer Physics Communications, 10,* 343-367.

Lavrentiev, M.M. (1967). *Some improperly posed problems of mathematical physics.* Springer-Verlag.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of tip, tongue, jaw and larynx movement. *Journal of the Acoustical Society of America, 50,* 1166-1179.

Lœvenbruck, H., & Perrier, P. (1993). Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands. In *Proceedings of the 3rd European conference on Speech communication and Technology,* Vol. 1 (pp.85-88). Berlin.

Maeda, S. (1979). An articulatory model based on statistical analysis. *Journal of the Acoustical Society of America, 65,* S1, S22(A).

Maeda, S. (1988). Improved articulatory model. *Journal of the Acoustical Society of America, 84,* Sup. 1, S146.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Modelling* (pp. 131-149). Kluwer: Academic Publishers.

Öhman, S.E. (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. *Journal of the Acoustical Society of America, 41(2),* 310-320.

Payan, Y., & Perrier, P. (1993). Vowel normalization by articulatory normalization : First attemps for vowel transitions. In *Proceedings of the 3rd European conference on Speech communication and Technology,* Vol. 1 (pp. 417-420). Berlin.

Perkel, J.S., Marc, H.C., Svirsky, M.A., Matthies, M.L., Garabieta, I., & Jackson, M.T.T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America, 6,* 3078-3096.

Perrier, P., Abry, C., & Keller, E. (1989). Vers une modélisation des mouvements du dos de la langue. *Journal d'Acoustique, 2,* 69-78.

Schroeder, M.R. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of the Acoustical Society of America, 41,* 1002–1010.

LE COMPILATEUR DE REGLES COMPOST ET SON APPLICATION POUR L'ANALYSE SYNTACTICO-PROSODIQUE DANS UN SYSTEME DE SYNTHESE PSOLA EN ALLEMAND

M. ALISSALI[2], S. FRENKENBERGER[1], M. KOMMENDA[3], B. SCHNABEL[1]
(noms en ordre alphabétique)

Centre National d'Etudes des Télécomunication (CNET), Lannion[1]
CNET et ENSSAT/IRISA, Lannion[2]
TU-Wien[3]

ABSTRACT

Development of high-quality text-to-speech (TTS) synthesis systems may be facilitated by the availibility of powerful programming tools allowing to process of complex linguistic informations. This paper presents the rule interpreter COMPOST which forms the kernel of such a tool.

COMPOST rules follow the notations classically used in phonetics and phonology. The rules operate on tree-structured objects which collect hierarchical attributes and are in accordance with the principles of object-oriented programming. This permits the easy development of modules for high-level tasks, such as prosody generation, involving the processing of concurrent informations at various linguistic levels. In the German version of the multilingual CNET PSOLA TTS system we start from a minimal syntactic analysis and subsequently carry out a much more refined prosodic analysis. The latter is based on rules which jointly exploit syntactical, morphological and phonological constraints.

1 INTRODUCTION

Inspirés par le formalisme des langages spécialisés et des environnements de programmation, certains outils ont été conçus pour la description des différentes transformations linguistiques qui conduisent d'un texte orthographique au signal sonore. Ils traitent ainsi non seulement des chaînes linéaires mais aussi des structures plus complexes (HERTZ 88, CARLSON & GRANSTRÖM 76, FRENKENBERGER et al. 91). Ceci permet d'exploiter des connaissances linguistiques beaucoup plus riches et de traiter une description du texte d'entrée contenant toutes les informations disponibles aux divers niveaux linguistiques (morphologie, syntaxe etc.)

Dans cette optique, le langage COMPOST a été conçu pour manipuler de tels objets complexes, organisés en une structure de travail arborescente. Suivant les principes élémentaires de la programmation orientée-objet, ces objets sont des instances d'objets génériques, définis au préalable selon les besoins particuliers d'un certain système de synthèse. Toutefois la topologie de ces objets est restreinte à des arborescences. Cette restriction permet une notation parenthésée de l'arbre et, par ce moyen, une écriture aisée des règles.

L'intelligibilité et l'agrément d'une voix synthétique dépendent largement d'un traitement approprié des phénomènes prosodiques. Les valeurs physiques de la fréquence fondamentale, des durées (segmentales et des pauses) et de l'intensité du signal de parole varient en fonction de la structure accentuelle, du rythme et du type d'intonation des différentes parties d'une phrase. Or, la détermination de ces faits prosodiques à partir d'un texte d'entrée fait appel non seulement à une analyse syntaxique de ce texte, mais aussi à l'exploitation des autres sources d'information disponibles, surtout morphologiques et phonologiques.

Dans la version en Allemand du système de synthèse PSOLA du CNET (BIGORGNE et al. 93), nous avons tenté de mettre en oeuvre une analyse syntaxique minimale. Le modèle prosodique est basé sur les seules unitées syntaxiques et sur le nombre de syllabes qu'elles contiennent.

Le langage COMPOST, tel qu'il est décrit dans le chapitre suivant, se prête bien au développement d'une telle analyse, comme nous le démontrerons dans la troisième partie de cet article.

2 COMPOST

2.1 Les bases de COMPOST

Le langage de programmation par règles de réécriture COMPOST (ALISSALI & BAILLY 93) est conçu d'après le formalisme couramment employé en linguistique (CHOMSKY & HALLE 68) :

$$C \rightarrow R / G + D \qquad (1)$$

Dans cette règle, C (la Cible) se réécrit R (Remplacement) si elle est précédée par G (le Contexte Gauche) et suivie par D (le Contexte Droit).

Dans le cas le plus simple, les composants de la règle (Cible, Remplacement et Contextes) sont des suites d'objets : entités élémentaires de traitement définies par l'utilisateur. Un objet est caractérisé par un nom unique et une liste d'attributs de type ordonné (entier, réel, caractère, booléen,...) et de type tri-valué ('trait').

L'objet suivant, nommé 'MOT', a été défini pour décrire les mots de l'allemand :

```
objet (MOT)
    trait -Adj, -Art, -Pron, -Prep, -Verb, -Nom, Poss, -EtOu ...
    entier Syll; Acc = 10;                                        (2)
```

La liste d'attributs comprend un ensemble de traits qui

serviront à déterminer la catégorie grammaticale du mot (Adj, Article, Pron, ...). Les attributs de type entier permettent des opérations de comptage comme, par exemple, le nombre de syllabes dans le mot (Syll). Dans une règle COMPOST, la mise en correspondance de la cible et des contextes peut être étendue aux tests sur les attributs. La syntaxe d'un tel test est :

<obj; att_ord op val_ord, ... [+-]att_trait,...> (3)

Dans cette règle, la valeur actuelle de l'attribut *att_ord* est comparée à la valeur *val_ord* par l'intermédiaire de l'opérateur *op* (==, <, >, ...). Dans le cas des traits, le test consiste à précéder l'attribut (*att_trait*) de la valeur testée ('+' ou '-'). Par exemple, avec la définition (2), l'écriture :

<MOT; +Verb> (4)

désigne un objet MOT si et seulement si il est marqué comme verbe (valeur '+' du trait 'Verb').
De même, les valeurs des attributs peuvent être modifiées dans la partie Remplacement d'une règle, c.à.d. lors de la génération de nouveaux objets (ou de la recopie d'objets existants, comme on le verra plus loin) :

<Obj; att1 = val1, ... [+-]att2,...> (5)

Pour les attributs de type "ordonné", l'opérateur d'affectation est '='. Pour les attributs de type "trait", la syntaxe est identique à celle du test, seul l'emplacement dans la règle (cible, contexte ou remplacement) détermine s'il s'agit d'un test ou d'une affectation. Par exemple, la règle suivante signifie qu'un mot qui se trouve entre un article et un nom , est un adjectif et non pas un verbe :

```
R1 = <MOT; +Verb, #M>
   -> <#M; -Verb, +Adj> / <MOT; +Art> + <MOT; +Nom>;(6)
```

Dans cette règle :
- la Cible est un objet 'MOT' marqué (+Verb)
- le Contexte Gauche est un objet 'MOT' marqué (+Art),
- le Contexte droit est un objet 'MOT' marqué (+Nom)
- le Remplacement est la recopie de l'entrée (#M) dans la sortie (#, cf. ci-dessous), ensuite les valeurs - et + seront affectées aux traits 'Verb' et 'Adj' respectivement.

Dans cet exemple, le remplacement consiste à modifier les attributs d'un objet existant, d'où la nécessité de maintenir une référence de cet objet. C'est le rôle de l'opérateur '#' dit de marquage/copie : l'objet 'MOT' mis en correspondance dans l'entrée est mémorisé grâce a la marque «#M». La même marque permet de récupérer cet objet dans la sortie et de modifier ensuite les valeurs de ses attributs.

2.2 Format Normalisé de COMPOST

Généralement, la complexité de la syntaxe d'un langage de programmation dépend de la puissance de celui-ci. Il faut alors faire un compromis entre la puissance de traitement et la simplicité de l'expression, qui, dans notre cas consiste à rester aussi proche que possible du formalisme (1). Pour ce faire, COMPOST définit une norme de la structure de données. Ce qui permet, d'une part, d'exprimer une relation d'arborescence avec une syntaxe simple, et d'autre part, de rendre implicite la tâche ardue de gestion de la structure de données.
La norme COMPOST définit une organisation en arbre n-aire dont chaque noeud est un objet. Une telle structure se révèle bien adaptée à certains traitements par règles de réécriture, tels que l'analyse syntactico-prosodique décrite au chapitre suivant.
Les relations hiérarchiques sont exprimées à l'aide de parenthèses : une parenthèse ouverte (resp. fermée) exprime le passage au niveau inférieur (resp. supérieur) comme dans le premier arbre de la Figure 1.

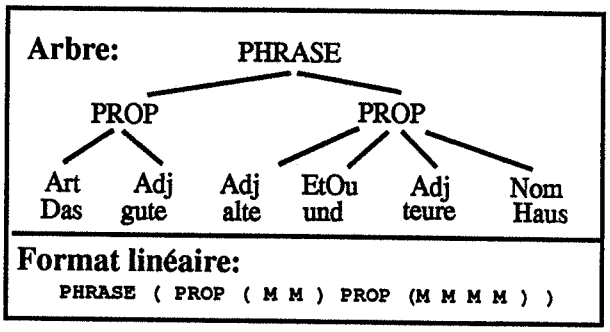


Figure 1: phrase "Das gute, alte und teure Haus" avant l'application des règles (7). Cet arbre peut être écrit par une forme linéaire ("Art Das" indique un objet Mot avec des attributs trait +Art, qui correspond au mot d'entrée "Das").

La règle suivante est un exemple de règle de modification d'arborescence. Elle se réfère aux deuxième et troisième niveaux de l'arbre précédent:

```
ChercheEnum =
      <MOT; +Adj, #w1> ) <PROP; +AfterComma> (
      <MOT; +Adj, #w2> <MOT; +EtOu, #w3>
      <MOT; +Adj, #w4> <MOT; +Nom, #w5>
   -> #w1 #w2 #w4; #w5
   / <MOT; +Art> + ;                          (7)
```

L'application de cette règle produit l'arbre de la Figure 2, où l'on voit disparaître le premier objet <PROP> et l'objet 'MOT' marqué par '+EtOu'.

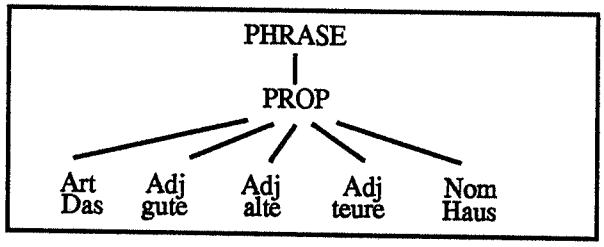


Figure 2: phrase "Das gute, alte und teure Haus" après l'application des règles

2.3 Les grammaires

Plusieurs règles peuvent être regroupées, dans un module appelé "grammaire", pour effectuer une opération précise de transduction sur un arbre d'objets. Une grammaire n'opère que sur un seul niveau de l'arbre d'entrée : les règles de la grammaire peuvent accéder en lecture à tous les niveaux de l'arbre, mais ne peuvent modifier que ceux du niveau déclaré dans l'entête de la grammaire.

Par conséquent, la recherche des cibles à traiter n'est effectuée que sur ce niveau. Ainsi, après le traitement (consommation) d'une cible, tous les objets se trouvant à sa droite et n'appartenant pas au niveau concerné sont recopiés tels quels dans la sortie. La recopie s'arrête dès que l'on rencontre un autre objet du niveau en question, cet objet devient le début de la nouvelle cible.
Pour chaque cible, les règles de la grammaire courante sont examinées dans leur ordre de déclaration. La première règle applicable, s'il y en a, est exécutée. Sinon (si aucune règle n'est applicable) l'objet courant est recopié sur la sortie. Dans les deux cas (application de règle ou simple copie), la cible courante est consommée par la règle : le traitement se poursuit avec la cible suivante.
La grammaire suivante, extraite de l'analyseur syntactico-prosodique de l'allemand (voir Chap. 3) met en œuvre des opérations complexes de transduction d'arborescences.

```
grammaire ConcConst [CONSTIT -> CONSTIT]
    ConcConst1 =
        <CONSTIT; #w1, Syll <= 2> (#A)
        <CONSTIT; +incert, #w2> (#B)
    -> <#w1; Syll = Syll + #w2.Syll> ( #A #B );          (8)
  ConcConst2 =
        <CONSTIT; +incert, #w1> (#A)
        <CONSTIT; #w2, Syll <= 7 - #w1.Syll> (#B)
    -> <#w2; Syll = Syll + #w1.Syll> (#A #B);            (9)
fingrammaire
```

Alors que tout objet traité par la règle (7) est explicitement mentionné dans la cible, dans la règle (8) les sous-arbres sont marqués par l'opérateur (#) et affectés par la règle (par exemple (#A #B) dans le remplacement). Ceci permet la fusion de deux sous-arbres, donc l'effacement de la racine de l'un en l'attachant à la racine de l'autre.

Dans le cas de l'exemple (règle (8)) cette fusion dépend, entre autres, du nombre de syllabes déterminé par l'attribut Syll de type entier. Cet attribut peut être testé, comme dans (3), pour vérifier que le nombre de syllabes ne dépasse pas une limite donnée :<CONSTIT; Syll<=2>.
Lors de la fusion, le nombre de syllabes des deux objets doit être additionné. Cela implique de référencer un objet autre que l'objet courant, comme exprimé par #w2.Syll dans la règle (8). Ainsi la valeur de l'attribut Syll est mise à jour à chaque nouvelle application d'une telle règle.
Les éléments de base et les outils nécessaires pour l'application que nous avons réalisée étant présentés, le chapitre suivant montre comment l'analyseur syntactico-prosodique de l'allemand peut être formalisé et réalisé en utilisant COMPOST.

3. L'ANALYSE SYNTACTICO-PROSODIQUE EN ALLEMAND

Dans ce chapitre, nous allons présenter l'implémentation de l'analyse syntactico-prosodique avec COMPOST dans la nouvelle version allemande du système de synthèse multilingue PSOLA du CNET (BIGORGNE et al. 93)

3.1 Objectif et point de départ de l'analyse syntactico-prosodique

3.1.1 Objectif

Tandis qu'une grammaire appliquée à l'acquisition d'une langue vivante vise à construire des phrases correctes (et doit omettre les phrases incorrectes) et à être complète (c.à.d. à couvrir tous les cas, même les plus rares), le but de l'analyse syntactico-prosodique dans un système de synthèse à partir du texte est de fournir UNE réalisation prosodique et rythmique d'une phrase qui soit adéquate (KOHLER 88). Le traitement des cas rares (même s'ils sont syntaxiquement corrects) ne joue pas un rôle décisif en synthèse vocale, et peut être en partie négligé.
Néanmoins, pour obtenir un système de synthèse de la parole d'une qualité satisfaisante, une modélisation appropriée des phénomènes prosodiques doit être rendue possible par l'analyse syntaxique. En allemand, ceci nécessite une bonne présentation des variations de F0 et des valeurs de durée pour chaque unité segmentale ainsi qu'une bonne prédiction de la durée des pauses.
La détermination de ces manifestations acoustiques de la prosodie exige tout d'abord une analyse à un niveau abstrait, guidée par la fonction linguistique de la prosodie.
En général, toute analyse syntactico-prosodique dans un système de synthèse s'appuie, dans un premier temps, sur les catégories grammaticales des mots et sur la ponctuation pour dériver la structure syntaxique d'un texte, et, de là, sa structure prosodique (HUBER et al. 87).
Dans ce chapitre, nous présentons une autre approche:
le point de départ est la reproduction de la structure de surface dans une structure arborescente à deux niveaux. Le niveau inférieur est constitué pour la chaîne de mots étiquetés (objets munis d'attributs) (voir 3.1.2) et le niveau supérieur pour des propositions (voir Figure 1).
L'analyse syntactico-prosodique elle-même se déroule en deux étapes, chacune insérant un nouveau niveau dans l'arbre:
étape1: découpage syntaxique en unités minimales et
étape 2: dérivation des constituants prosodiques.

3.1.2 Etiquetage et segmentation en proposition

Notre analyse syntactico-prosodique utilise d'une part la structure de surface du texte d'entrée, dont chaque mot est étiqueté en fonction de son appartenance à une ou plusieurs classes de mots (noms communs, mots fonctionnels, adjectif et verbe), et d'autre part un premier découpage de la phrase d'entrée en propositions à l'aide des signes de ponctuation.
Les classes de mots:
* Les noms communs sont correctement détectés grâce à leur majuscule initiale.
*A l'aide d'un répertoire, les mots fonctionnels sont classés en 17 classes ayant souvent plusieurs étiquettes (par exemple, "meine" = verbe, pronom personnel).
* Actuellement, beaucoup d'analyseurs syntaxiques de l'Allemand se fondent sur la détection du verbe conjugué (SCHNABEL & ROTH 90). Mais, comme cette détection est

difficile, elle est source d'erreurs dans les systèmes de synthèse. Aussi avons-nous élaboré une approche qui reste la plus indépendante possible de la classification du verbe.
Ainsi, si une forme étiquetée comme verbe (par la morphologie) peut être un adjectif ou un verbe ("Die züchtigen [adjectif] Eltern züchtigen [verbe]"):
- nous la considérons toujours comme un adjectif, même dans les contextes ambigus (c.à d. ceux où elle pourrait être un verbe ("Der verdient [verbe] Geld", "Das verdiente [adjectif] Geld", où dans le premier cas, "verdiente" serait faussement étiqueté comme adjectif), nous acceptons donc un étiquetage erroné dans des tels cas.
- nous ne considérons cette forme ambiguë comme étant un verbe conjugué que si les contextes indiquent sans ambiguité qu'elle ne peut pas être un adjectif ("Der Mann verdiente [verbe] viel Geld").
Ainsi, la règle 7, telle qu'elle est définie dans le chapitre 2, transforme donc les attributs (+Verb/+Adj) en attributs (+Adj /-Verb) dans des séquences fréquentes (+Art +Adj +Nom), comme "die züchtigen Eltern". Par contre, dans la deuxième occurence de 'züchtigen" l'amiguïté n'est pas levée et on garde l'attribut (+Verb).

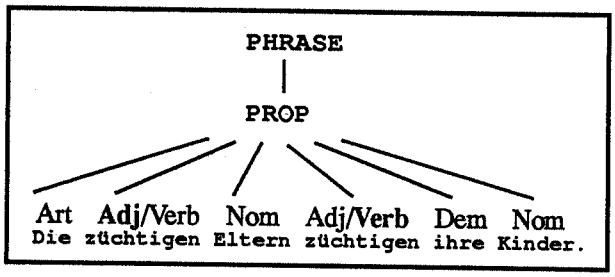


Figure 3: l'étiquette de la première occurrence de mot "züchtigen" est regardée comme Adj

Les propositions:
* Les signes de ponctuation sont traités comme des marqueurs de fin de propositions. Une proposition (PROP) est considérée comme l'objet du niveau supérieur qui gouverne les mots, donc les objets du niveau inférieur. Nous distinguons deux sortes de propositions, progrédiente et terminale, d'après le genre de signe de ponctuation.
Mais un signe de ponctuation n'indique pas toujours une frontière syntaxique, comme dans le cas d'une énumération (par exemple: "Das gute, alte und teure Haus ....") La structure de cette énumération, après l'exploitation des signes de ponctuation est représentée sur la Figure 1.
L'application de la règle (7) (expliquée dans le chapitre 2) permet de modifier la structure arborescente de façon à ce que l'analyse syntaxique puisse détecter correctement un seul groupe par la suite. La structure résultant de l'application de la règle (7) est représentée sur Figure 2.

3.2 Découpage de la proposition en unités syntaxiques et prosodiques

Il est possible d'identifier, au moins en Allemand, des unités syntaxiques minimales, qui, en règle générale, ne seront jamais décomposées en unitées prosodiques de taille inférieure. L'exemple le plus fréquent est celui des groupes nominaux comme "das alte Haus" (la vieille maison) ou "mit seinen Freunden" (avec ses amis). Une fois que ces unités minimales sont déterminées, elles peuvent être rassemblées en constituants prosodiques, en fonction du contexte et des critères rythmiques, entre autres.
Nous avons donc défini que:
- un groupe (GROUP) représente l'objet au premier niveau intermédiaire de la structure arborescente entre MOT et PROP, correspondant à l'unité syntaxique minimale. Il domine un mot ou une suite de mots étiquetés et syntaxiquement liés (voir Figure 4).
- un constituant (CONSTIT) représente l'objet au deuxième niveau intermédiaire de la structure arborescente entre GROUP et PROP. Dérivé par concaténation ou transformation de la succession de groupes, qu'il domine par la suite, il correspond à l'unité prosodique. Donc en vue de sa réalisation orale, il est marqué par l'intonation, l'accentuation, et l'éventuelle pause finale.
Dans les deux chapitres suivants, nous décrivons d'abord la détermination des groupes syntaxiques, puis leurs transformation en constituants prosodiques.

3.2.1 Analyse syntaxique
La détermination des groupes se base sur la succession de classes de mots, et part du principe qu'un groupe est le plus souvent introduit par un mot fonctionnel et terminé par un nom commun (SORIN et al. 87).
Cette règle sert de base à des règles plus complexes qui prennent en compte le nombre et la classe des mots entourés par le mot fonctionnel et le nom commun (adjectif, adverbe ...). De plus, des groupes ne contenant aucun nom commun, tel que groupe verbal (GroupVerb) et groupe adverbial (GroupAdv), ont été répertoriés par d'autres règles.
Les groupes les plus courants (mis en italique) sont:

GroupNomPrepCourt	Geschichten *von Goethe*
GroupNomPrep	*mit seinem neuen Fahrrad*
GroupNomExp	*die* von der abendlichen Sonne *beleuchteten Granitfelsen...*
GroupNom	*Die züchtigen Eltern*
GroupInitSubord	... gekommen, *weil* sie ...
GroupFinInf	sinnvoll *zu verwenden sein*
GroupVerb	... *züchtigen* ...
GroupAdv	... sah das Haus *allein*...

Au total, nous avons déterminé une cinquantaine de groupes, chacun généré par une règle de grammaire. Les règles contenues dans la grammaire sont écrites de façon hiérarchique. Plus une règle est importante, plus elle figure haut dans la grammaire.
L'accentuation du groupe pose un problème particulier (ADRIAENS 84). Le défi est de positionner au moins un accent par groupe sans créer de l'emphase, ce qui serait le cas par exemple en accentuant un mot fonctionnel. Parmi les mots lexicaux, nous cherchons plusieurs candidats susceptibles de recevoir un accent. Contrairement à d'autres approches (FRENKENBERGER & KOMMENDA 88), nous ne recherchons pas un mot seul, mais nous associons une

"valeur accentuelle" (VALEUR) à chaque mot, valeur qui, dans un traitement ultérieur, définira le degré d'accentuation sur une échelle de 10 à 60. On tient compte de cette valeur pour l'accent: plus la valeur est basse, moins la probabilité d'y mettre un accent sera grande.
Par exemple, dans un groupe nominal, le nom commun et l'adjectif qui le précède ont la même valeur accentuelle, qui est la plus grande du groupe, ce qui permet un traitement neutre de la prosodie.
Dans l'implémentation, les VALEURs sont représentées par un attribut de type entier des objets mots. La règle 13 implémente le cas du groupe nominal ; elle attribue la VALEUR 10 au déterminant (Art), la VALEUR 50 à l'adjectif (Adj), et la VALEUR 50 au nom commun (Nom).

```
GroupeNominal =
  <word; +Art, #w1> <word; +Adj, #w2><word; +Nom, #w3>
    -> GroupNom ( <#w1; VALEUR =10><#w2; VALEUR =50>
    <#w2; VALEUR =50> ) ;                                (10)
```

En appliquant la règle 13 à l'exemple de la Figure 3, on obtient le résultat donné Figure 4.
A partir de cette représentation de la structure de surface, nous pouvons procéder à l'analyse prosodique.

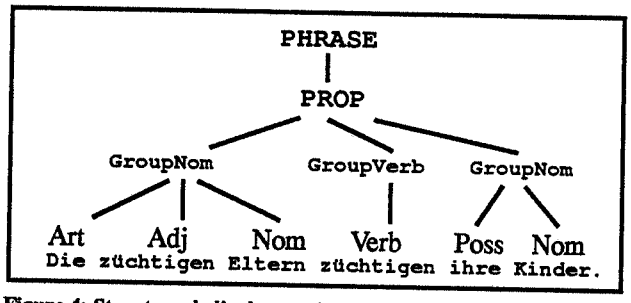


Figure 4: Structure de l'arbre après analyse syntaxique

3.2.2 Découpage prosodique

Contrairement aux groupes qui sont le résultat d'une analyse purement syntaxique, les constituants (prosodiques) sont issus de l'application de restrictions d'une part rythmiques (le nombre de syllabes) et d'autre part syntaxiques (unités minimales).
Etant donné qu'un constituant n'est pas une unité précisément définie, et qu'il existe des variations interlocuteurs (voire intralocuteurs) importantes, nous avons considéré que chaque constituant:
- ne doit pas couper les groupes en deux,
- dépend du nombre de syllabes qu'il contient.

De même que les règles qui déterminent les groupes se basent sur la succession des classes de mots, la détermination des constituants se base sur la succession des groupes; leur nature progrédient (+prog) ou terminal (+term) est conditionné par les signes de ponctuation.
Pour ce faire, on passe par les étapes suivantes :
- toute suite de mots comprise entre deux signes de ponctuation (sauf pour l'énumération) est considérée comme un constituant.
- les constituants obtenus lors de la première étape sont ensuite analysés afin de les découper en trois types de constituants :
  - les constituants certains,
  - les constituants verbaux, et
  - les constituants incertains, constituants ne peuvant pas être classés dans les deux premiers types.
- on concatène deux ou plusieurs constituants incertains adjacents dans un seul constituant.

La premier étape consiste à insérer un nouveau niveau (intermédiaire deux) dans la structure aborescente où on crée pour chaque proposition un nouveau objet (CONSTIT) ayant les mêmes attributs (+prog, +term).
La deuxième étape effectue le découpage du constituant en plusieurs sous-constituants sur le même niveau de la structure arborescente. Pour déterminer les différents types de constituants, on applique des règles dont la structure est la suivante:

```
<CONSTIT; #ph1>
            ( <GroupNom; #g1>
    -> <CONSTIT; +prog> ( #g1 ) #ph1 (
        / + GroupVerb ;                                  (11)
```

Un constituant issu de la première étape est divisé en un constituant progrédient séparé du constituant terminal, si le groupe nominal est suivi par un groupe verbal.
La Figure 5 représente l'application de la règle (11) sur l'exemple de la Figure 4: le groupe nominal "Die züchtigen Eltern" est étiqueté comme constituant progrédient, car il est suivi par un verbe, alors que le groupe verbal "züchtigen" et le groupe nominal terminal "ihre Kinder" sont conservés comme constituant terminal.

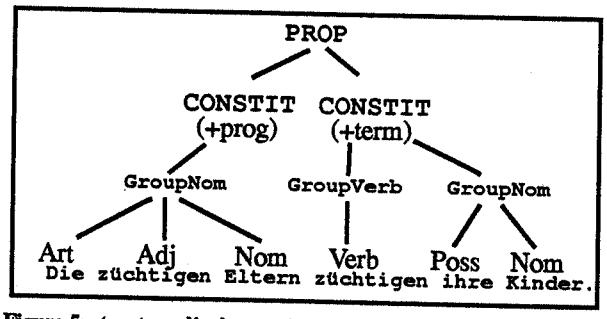


Figure 5: structure d'arbre après l'application de règles, qui traitent des constituants.

Pendant que l'étape 2 effectue la subdivision d'un constituant initial en plusieurs sous-constituants, l'étape 3 concatène plusieurs constituants en un seul constituant (unité prosodique maximale).
Certains constituants comme le groupe nominal au génitif, le groupe final d'une énumération ou le groupe adverbial nécessitent un étiquetage comme constituants "incertains", car ils sont étroitement liés au constituant précédant ou au suivant. Qu'un tel constituant soit finalement considéré comme unité prosodiquement autonome, dépend des conditions syntactico-sémantiques (a priori inaccessibles dans le cadre d'une analyse de surface) et des restrictions rythmiques (nombre de syllabes du constituant lui-même).
Le regroupement de plusieurs constituants nécessite la concaténation des branches du sous-arbre, comme nous

l'avons vu dans le chapitre 2. Dans un premier temps, deux constituants adjacents sont rassemblés si:
- le constituant précédant a moins de 4 syllabes
- le constituant incertain a moins de 3 syllabes
- la somme des syllabes du constituant incertain et de celles du constituant suivant est inferieure à 7,

correspondant aux règles (8) à (9) du Chapitre 2.
Par la suite, dans toute suite de trois constituant se succédant, deux sont concaténés afin d'éviter qu'un constituant trop court ne reste comme unité autonome.
Une fois les constituants établis, c.à.d. que les marqueurs prosodiques sont positionnés dans la phrase et que chaque mot a reçu une VALEUR d'accentuation, toutes les données nécessaires peuvent être transmises au module suivant qui effectue le calcul des variations de F0 et des valeurs de durée segmentale et des pauses

3.3 Résultats

L'analyse syntactico-prosodique décrite en 3.2, qui se base essentiellement sur la segmentation de la phrase en unités syntaxiques minimales et sur des critères rythmiques, permet d'obtenir des résultats tout à fait prometteurs.
Si l'atout de l'analyse syntaxique ne réside pas dans sa complexité, elle gagne par contre en robustesse (c.à.d. en indépendance vis-à-vis du texte d'entrée), en assurant un découpage correct pour presque tous les types de phrases.
L'application de plusieurs boucles de concaténation de groupes ou de constituants en un constituant autonome permet en particulier d'éviter des constituants trop courts.
Nous avons testé nos règles en effectuant un découpage syntactico-prosodique sur un texte littéraire de BORGES de 175 phrases longues. Il va de soi que nous n'avons pas utilisé ce texte pour le développement des règles.
Le système a trouvé 814 marqueurs de constituants, dont 175 grâce aux points (marqueur terminal) et 190 grâce aux virgules. Des 449 marqueurs restants, 17 (soit 4%) sont incorrectement positionnés. Parmis ces erreurs nous n'avons observé qu'un seul cas où une phrase longue (ne contenant aucune virgule) n'a reçu aucun marqueur:
"..., dessen monströse Anatomie sich mir solchermaßen indirekt wie durch ihren Schatten als die eines Tieres oder eines Gottes enthüllte."
Les autres erreurs sont dues à un mauvais positionnement du marqueur. Nous avons rencontré un cas où un groupe était mal déterminé à cause d'un adjectif en majuscule:
"Der Verkauf *des Roten # Hauses* und dieser..."
Nous avons également certains cas où la concaténation de constituants a entraîné des erreurs. Ici un exemple où le mot fonctionnel a été mal traité:
"*Seiner unmaßgeblichen Meinung # nach* sei der Señor..."
Plusieurs de ces cas resteront insolubles: ils peuvent être dus à la sémantique ou à des particularités orthographiques, comme des adjectifs en majuscules.
Certaines de ces erreurs par contre, peuvent être corrigées en ajoutant des règles de concaténation dans la détermination des CONSTIT ou en corrigeant certaines entrées dans la liste des mots fonctionnels.

4 CONCLUSION

Cet article présente le compilateur de règles COMPOST et son application à l'implémentation d'une analyse un syntactico-prosodique en allemand. La possibilité qu'offre le langage COMPOST de définir des objets complexes, possédant des listes d'attributs de type ordonné et tri-valué, et leur représentation dans une structure arborescente sont largement exploitées par cette application.Ainsi, la structure d'une phrase allemande a été reproduite en arbre.
L'ensemble des règles de découpage syntactico-prosodique, implémenté avec COMPOST, permet de traiter tout texte allemand avec une assez grande fiabilité et une grande robustesse: pour n'importe quelle phrase allemande le système propose UN découpage prosodique acceptable.
Le respect des unités syntaxiques minimales et l'affectation attentive de valeurs accentuelles à chaque mot garantissent une distorsion minimale de la structure syntactico-prosodique dans les quelques découpages erronnés.

5 REFERENCES

ADRIAENS L.M.H. 84 ***"A Preliminary Description of German Intonation"*** IPO (Inst. voor Perceptie Onderzoek), Annual Progress report 19, p.36-45, Eindhoven, 1984.

ALISSALI, M. & G. BAILLY 1993: *"COMPOST: A Client-Server Model for Applications Using Text-to-Speech"*. EuroSpeech'93, Berlin, Vol. 3, p. 2095-2098.

BIGORGNE D., O. BOEFFARD, B. CHERBONNEL, F. EMERAD, D. LARREUR, J. L. LE SAINT-MILON, I. METAYER, C. SORIN, S. WHITE 1993: ***"Multilingual PSOLA Text-to-Speech System"***. ICASSP'93, April 27-30, 1993, Minneapolis (Minnesota, USA).

CARLSON R. & B. GRANSTRÖM 76: ***"A Text-to-Speech System Based Entirely on Rules."*** In IEEE International Conference on Acoustics, Speech, and Signal Processing, p.789-688, Philadelphia, 1967.

CHOMSKY N. & M. HALLE 68:*"The Sound Pattern of English"* Harper & Row, New York, 1968.

FRENKENBERGER S., & M. KOMMENDA 88: ***"Prosodiebestimmung in der Sprachsynthese"*** In Proc. 4th österreichische Artificial Intelligence Tagung, Informatik-Fachberichte, vol. 176, p. 62-71, Berlin New York Tokyo, 1988.

FRENKENBERGER, S., M. KOMMENDA, S. MOSSMÜLLER 91: ***"A Multi-Linear Representation and Rule Formalism for Phonetics in Text-to-Speech Synthesis."*** In Proceedings of the XII'th international Congress of Phonetic Sciences, vol. 2, p. 518-521, Aix-en-Provence, 1991.

HERTZ, S.R. 88: ***"The Delta Programming language: an integrated approach to non-linear phonology, phonetics and speech synthesis."*** Working Papers of the Cornell Phonetic Lab, Nr.2, p. 69-122, 1988.

HUBER K. & AL. 87:*"Sprachsynthese ab Text."* In: Tillman H. G. and G. Willée (ed.): ***"Analyse und Sythese gesprochener Sprache."*** Georg Olms Verlag, Hildesheim, p.26-33, 1987.

KOHLER K. 88: ***"An Intonation Model for a German Text-to-Speech System"*** Proceedings of the 7th FASE Symposium - SPEECH'88, Edingburgh, vol. 4, p. 1241-1247, 1988.

SCHNABEL, B. & H. ROTH 1990: ***"Automatic linguistic processing in a german Text-to-Speech synthesis System"***. ESCA/ETRW on Speech Synthesis, AUTRANS (France), p.121-124, 1990.

SORIN, C., D. LARREUR, R. LLORCA 87: ***"Rhythm-based prosodic parser for text-to-speech systems in French."*** Procedings of the 11th Int. Cong. of Phonetic Sciences, Tallin, Estonie, URSS, vol.1, 125-128, 1987.

Comparaison entre la variabilité des formants et la variabilité d'estimations des cibles formantiques

Michel PITERMANN†* et Jean CAELEN‡

†Institut des Langues Vivantes et de Phonétique, Université Libre de Bruxelles, CP 110, 50 av. F. D. Roosevelt, 1050 Bruxelles, Belgique

‡Institut de la Communication Parlée, 46 av. Félix Viallet, 38031 Grenoble CEDEX, France

Résumé

La théorie des cibles en production de la parole postule que les voyelles sont caractérisées par des valeurs formantiques appelées *cibles*. Ces cibles seraient indépendantes de la vitesse d'élocution et du degré d'accentuation de la voyelle. Une mise à l'épreuve expérimentale de cette hypothèse est présentée dans ce document. La variabilité d'estimations de cibles a été comparée à la variabilité des formants d'un [a] dans un contexte [iai] pour des vitesses d'élocution et des degrés d'accentuation différents. Les cibles non atteintes ont été estimées au moyen d'une modélisation dynamique linéaire et d'une modélisation cinématique non linéaire des trajectoires formantiques. En dessous d'une certaine vitesse d'élocution, nos résultats étaient en accord avec l'hypothèse selon laquelle la voyelle [a] est caractérisée par des cibles formantiques invariantes par rapport à la vitesse d'élocution et au degré d'accentuation. De plus, un paramètre dynamique du modèle non linéaire était indépendant de la vitesse d'élocution et permettait de distinguer le [a] accentué du [a] non accentué.

Mots-clés : *cible non atteinte, invariance, réduction vocalique, modélisation dynamique, formant.*

1 Introduction

Des expériences ont montré que les voyelles soutenues peuvent être identifiées à l'aide de leurs deux ou trois premiers formants [Helmholtz 1954; Joos 1948; Delattre et al. 1952]. En parole continue, les fréquences formantiques d'une voyelle varient lorsqu'elle est produite par un même locuteur pour un même contexte phonétique, mais à des vitesses d'élocution et des degrés d'accentuation différents [Lindblom 1963]. Cependant, la plupart des théories en production de la parole supposent explicitement ou implicitement que les voyelles sont caractérisées par des valeurs formantiques appelées *cibles* qui seraient indépendantes de la vitesse d'élocution et du degré d'accentuation de la voyelle. Bien qu'il s'agisse d'une hypothèse centrale en production et en perception de la parole, l'existence de cibles formantiques invariantes n'a pas encore reçu à l'heure actuelle de confirmation expérimentale satisfaisante. Le travail présenté dans ce document vise à combler cette lacune.

Pour comparer la variabilité des cibles à la variabilité des formants, un moyen d'estimer les cibles non atteintes est nécessaire. L'approche que nous avons choisie consiste à modéliser cinématiquement ou dynamiquement des transitions formantiques [Caelen 1985; Beautemps 1993]. Dans le premier cas, la cible est définie par l'asymptote de la courbe, dans le second cas, elle est définie par la solution stationnaire du modèle dynamique. Malheureusement, la variabilité des estimations de cibles a toujours été aussi importante que la variabilité des formants. La sensibilité du résultat au bruit peut en être la cause [Pitermann and Caelen 1993]. Par conséquent, nous avons évalué la robustesse au bruit de nos estimateurs dans le but de contrôler l'impact du bruit sur les estimations, puis nous avons développé des techniques pour accroître la précision des résultats.

2 Méthode

Un peu plus de 800 transitions des deux premiers formants d'un [a] en contexte [iai] ont été analysées. Le [a] a été accentué ou non accentué, et le syntagme porteur [il*iai*media] a été produit par un locuteur masculin à 10 vitesses d'élocution différentes contrôlées par métronome. Le locuteur synchronisait le premier [i] du syntagme porteur avec une impulsion du métronome et le [d] avec l'impulsion suivante. Les 10 rythmes furent choisis entre 54 et 152 battements par minute. Ces deux extrêmes correspondaient à une vitesse d'élocution anormalement lente et au dé-

* Boursier de la fondation David et Alice Van Buuren, Belgique

bit maximum que pouvait contrôler le locuteur. Pour chacun des 20 couples (débit, accent), le syntagme a été répété jusqu'à ce qu'au moins 20 exemples aient été produits correctement, et nous avons pris pour hypothèse que les 20 exemples étaient caractérisés par la même dynamique.

Chaque transition formantique comprise entre le [i] et le [a] suivant a été analysée au moyen de deux modélisations unidimensionnelles: l'ajustement sigmoïdal qui est un modèle cinématique non linéaire, et une classe de modèles dynamiques linéaires.

L'équation de la sigmoïde était:

$$x(t) = c + \frac{a}{1 + e^{-b(t-t_0)}} \qquad (1)$$

où c et $a + c$ désignaient ses deux asymptotes, b son coefficient de relaxation, t le temps, et t_0 l'intervalle de temps séparant le point d'inflexion de la courbe et l'origine temporelle (cf. Fig. 1 pour l'illustration). L'estimation de la cible était donc $x^* = a + c$. La sigmoïde était ajustée sur des transitions formantiques comprises entre le dernier point du plateau du [i] et l'extremum suivant, c'est-à-dire un maximum pour F_1 et un minimum pour F_2. Pour chaque couple (débit, accent), les paramètres du modèle ont été estimés en utilisant simultanément les 20 transitions correspondantes grâce à l'algorithme de Viterbi. Le principe était de déterminer les 3 paramètres structuraux (a, b, c) par minimisation de l'erreur totale de modélisation en ayant fixé les valeurs des 20 paramètres t_0, puis d'utiliser les 3 valeurs déterminées pour estimer les 20 paramètres t_0 par une minimisation des 20 erreurs de modélisation ou par programmation dynamique; le processus était réitéré jusqu'à ce qu'il converge. Les erreurs de modélisation étaient minimisées par l'algorithme de Levenberg-Marquardt [Press et al. 1988]. Utiliser simultanément 20 transitions pour estimer des paramètres structuraux accroît la précision du résultat.

L'équation de la classe de modèles linéaires choisie était:

$$x_n = a_0 + \sum_{i=1}^{k} a_i x_{n-i} \qquad (2)$$

où x_j représentait la valeur du formant à l'instant j, k était l'ordre du modèle, et a_l en étaient les paramètres. L'estimation de la cible définie par la solution stationnaire du modèle était donc $x^* = a_0/(1 - \sum_{i=1}^{k} a_i)$. Les segments de transitions formantiques analysés étaient compris entre le point d'inflexion de la courbe déterminé grâce à l'ajustement sigmoïdal (paramètre t_0) et l'extremum suivant. Pour chaque couple (débit, accent), les paramètres du modèle étaient estimés par une régression linéaire utilisant simultanément les 20 segments de transitions correspondants. Tous les points $(x_n, x_{n-1}, \ldots, x_{n-k})$ de ces 20 segments étaient inclus dans la régression.

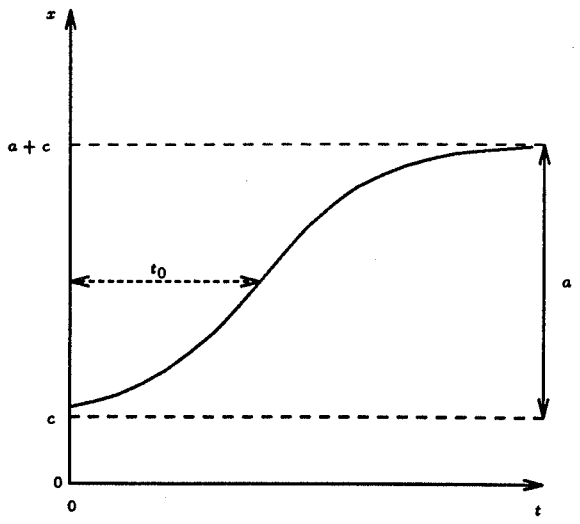


FIG. 1 - *Sigmoïde.*

Cette technique permet d'estimer les paramètres du modèle à l'aide de plusieurs transitions sans qu'il soit nécessaire de les synchroniser. Pour chaque couple (débit, accent), l'ordre du modèle était déterminé par une analyse des erreurs standards d'estimation des paramètres. Les paramètres et leurs erreurs standards étaient évalués pour différents ordres de modélisation, et l'ordre le plus élevé pour lequel tous les paramètres étaient significatifs à 95 % était retenu. Un paramètre était considéré significatif si sa statistique-t était supérieure à 2.

Les erreurs standards d'estimation des cibles ont été évaluées pour la modélisation linéaire par *bootstrap* [Efron and Tibshirani 1984]. Pour chaque couple (débit, accent), une fonction de distribution de l'estimation de la cible a été synthétisée par cette méthode. L'erreur standard était définie par la demi-largeur de l'intervalle laissant 16 % des données dans chaque queue de la distribution. L'intervalle considéré contenait donc les 68 % classiques des données.

Afin d'estimer une fréquence formantique et son erreur standard pour un couple (débit, accent), l'extremum de l'évolution temporelle du formant a été mesuré pour le [a] pour chaque transition du groupe (maximum de la courbe pour F_1, minimum pour F_2). La fréquence formantique du groupe était définie par la médiane des 20 mesures, son erreur standard par la demi-largeur de l'intervalle laissant 16 % des mesures dans chaque queue de la distribution.

Les trajectoires formantiques ont été calculées par une méthode combinant une analyse LPC et une analyse cepstrale dans une fenêtre de 13.5 ms de long se déplaçant par pas de 6.75 ms. Aucun filtrage temporel n'a été utilisé.

3 Résultats

L'ordre du modèle linéaire nécessaire pour décrire un groupe de 20 segments de transitions était toujours inférieur ou égal à 3.

Une comparaison entre les estimations de cibles et les fréquences formantiques pour des vitesses d'élocution et des degrés d'accentuation différents est présentée sur la Fig. 2. Les deux graphiques du dessus ont trait à F_1 et ceux du dessous à F_2, ceux de gauches sont relatifs au [a] accentué, et ceux de droite au [a] non accentué. Les fréquences formantiques sont représentées par une étoile (*), les estimations de cibles obtenues par modélisation linéaire par un rond (o), et celles obtenues par modélisation sigmoïdale par un losange ($\diamond$). Les erreurs standards sont indiquées par les traits verticaux. Les ordonnées sont graduée en Hz et les abscisses, qui représentent l'intervalle de temps séparant deux impulsions du métronome, sont graduées en ms. Les points correspondant aux débits rapides sont donc situés sur la gauche de chaque graphique.

L'estimation du coefficient de relaxation b de la sigmoïde en fonction de la vitesse d'élocution est représentée sur la Fig. 3. Le graphique de gauche a trait à F_1, celui de droite à F_2. Les étoiles (*) sont relatives au [a] accentué, les losanges ($\diamond$) au [a] non accentué. Les ordonnées sont graduées en unité arbitraire de dimension $(1/s)$, les abscisses, qui représentent à nouveau l'intervalle de temps séparant deux impulsions du métronome, sont graduées en ms.

Le coefficient de relaxation b de la sigmoïde était indépendant de la vitesse d'élocution pour les transitions du premier formant (cf. graphique de gauche de la Fig. 3). Les droites de régression linéaire obtenues pour le [a] accentué (les étoiles (*)) et pour le [a] non accentué (les losanges ($\diamond$)) étaient toutes les deux horizontales, c'est-à-dire que le paramètre correspondant à la pente était non significatif dans les deux cas, même lorsque l'intervalle de confiance classique de 95 % était réduit à 50 %. Les médianes des coefficients de relaxation pour le cas accentué et pour le cas non accentué valaient respectivement 0.38 ± 0.015 et 0.34 ± 0.035 (les erreurs mentionnées sont les demi-intervalles interquartiles).

4 Discussion

Les valeurs de F_1 correspondant aux 4 vitesses d'élocution les plus faibles pour le [a] accentué étaient à peu près égales à 590 Hz (cf. les 4 (*) de droite du graphique supérieur gauche de la Fig. 2). Par conséquent, si une cible existait dans ce contexte, elle devait valoir à peu près 590 Hz. Pour ces quatre points, la cible estimée par le modèle sigmoïdal était très proche de cette valeur, celle obtenue par modélisation linéaire était proche ou supérieure. La deuxième asymptote du modèle sigmoïdal était donc un estimateur de cible non biaisé, tandis que la solution asymptotique du modèle linéaire surestimait peut-être la cible.

La comparaison entre la variation du premier formant et la variation des estimations de la cible en fonction de la vitesse d'élocution indique que les estimations de la cible variaient moins que le formant pour les 7 débits lents du [a] accentué et les 6 débits lents du [a] non accentué. Pour ces débits, les estimations de la cible du [a] non accentué obtenues par modélisation linéaire valaient aussi à peu près 590 Hz. Bien que les résultats doivent être confirmés par l'analyse de données supplémentaires et par un traitement quantitatif, ils sont favorables à l'hypothèse selon laquelle le [a] serait caractérisé par une cible indépendante de la vitesse d'élocution et du degré d'accentuation.

Ces résultats soutiennent aussi l'hypothèse selon laquelle la dynamique est plus importante pour caractériser une voyelle que l'information contenue dans sa partie pseudostationnaire [Strange 1987; Cassidy and Harrington 1992], *même pour un contexte* $V_1 V_2$. De plus notre méthode indique quelle information pourrait être utilisée pour identifier les voyelles.

Au-delà d'une certaine vitesse d'élocution, les mesures formantiques et les estimations de la cible coïncidaient. Ceci suggère que pour ces vitesses d'élocution, l'information sur la cible non atteinte ne pouvait plus être extraite de la dynamique du formant par notre méthode. Des analyses plus approfondies pourraient déterminer si la brusque variation des estimations de la cible était le reflet d'un réel changement de production survenant lorsque la contrainte de débit devenait trop forte, et si pour ces débits, [iai] est toujours perçu.

Pour le second formant, les résultats sont nettement moins convaincants. Deux hypothèses pourraient expliquer ce phénomène. Premièrement, la classe de modèles linéaires et le modèle sigmoïdal pourraient décrire correctement le mouvement des articulateurs [Perrier et al. 1990]. Pour la production de [iai], le lien entre le mouvement des articulateurs et le premier formant est à peu près linéaire, ce qui n'est pas le cas pour le second formant [Majid et al. 1987]. Ce dernier devrait donc être modélisé moins linéairement que ce qui a été réalisé jusqu'à présent. Deuxièmement, l'affiliation entre le second formant et les résonances des cavités du conduit vocal pose problème avec la transition [ia], ce qui n'est pas le cas pour le premier formant [Bailly 1993]. Dans cette transition, une permutation entre la cavité affiliée à F_2 et la cavité affilié à F_3 ou à F_4 entraînerait une transition de F_2 obéissant à deux lois dynamiques qui se succéderaient. Remplacer le suivi des formants

par un suivi en résonances pourrait donc améliorer significativement les résultats relatifs à F_2.

Bien qu'une étude en cours sur l'erreur standard d'estimation du coefficient de relaxation de la sigmoïde devrait nous permettre d'affiner puis de confirmer les résultats présentés, ce paramètre dynamique était indépendant de la vitesse d'élocution pour le premier formant du [a]. De plus, il permettait de distinguer la voyelle accentuée de la voyelle non accentuée. Le coefficient de relaxation d'une sigmoïde étant un paramètre purement dynamique, nous avons un nouvel argument en faveur de l'importance prépondérante de la dynamique pour caractériser une voyelle par rapport à l'information contenue dans sa partie pseudostationnaire. De plus, notre méthode indique quel type d'information pourrait être utilisée pour mesurer l'accent de la voyelle.

Les résultats relatifs au second formant sont à nouveau moins bons que ceux du premier. Les hypothèses explicatives ont déjà été présentées.

5 Conclusion

La variabilité d'estimations de cibles a été comparée à la variabilité des fréquences formantiques d'un [a] en contexte [iai] pour des vitesses d'élocution et des degrés d'accentuation différents. Les cibles non atteintes ont été estimées par une modélisation dynamique linéaire et par un modèle cinématique sigmoïdal des trajectoires deux premiers formants.

En dessous d'une certaine vitesse d'élocution, la variabilité des estimations de la cible du premier formant était inférieure à la variabilité de ce formant. Bien que ce résultat doit être confirmé par l'analyse de données supplémentaires et par un traitement quantitatif, il constitue un argument en faveur de l'hypothèse selon laquelle la voyelle [a] est caractérisée par une cible formantique indépendante de la vitesse d'élocution et du degré d'accentuation.

Le coefficient de relaxation de la sigmoïde modélisant la transition du premier formant était indépendant de la vitesse d'élocution et permettait de distinguer la voyelle accentuée de la voyelle non accentuée.

Ces résultats soutiennent aussi l'hypothèse selon laquelle l'information dynamique caractérise mieux une voyelle que l'information contenue dans sa partie pseudostationnaire. De plus, notre méthode indique quel type d'information dynamique pourrait être utilisée pour caractériser et identifier une voyelle.

Références

Bailly, G. (1993, september). Resonances as possible representation of speech in the auditory-to-articulatory transform. In *Proceedings of Eurospeech '93*, Volume 2, Berlin, Germany, pp. 1511–1514. catalyst consult: ESCA.

Beautemps, D. (1993). *Récupération des gestes de la parole à partir de trajectoires formantiques: Identification de cibles vocaliques non-atteintes et modèles pour les profils sagittaux des consonnes fricatives.* Ph. D. thesis, INPG, Grenoble.

Caelen, J. (1985, june). Introduction à une segmentation cinématique. In *Actes des 14e Journées d'Étude sur la Parole*, École Nationale Supérieure des Télécommunications, Paris, France, pp. 129–132. GALF.

Cassidy, S. and J. Harrington (1992, december). Investigating the dynamic structure of vowels using neural networks. In *Proceedings of the Fourth Australian International Conference on Speech Science and Technology*, University of Queensland, Brisbane, Australia, pp. 495–500. Australian Speech Science and Technology Association.

Delattre, P., A. M. Liberman, F. S. Cooper, and L. J. Gerstman (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word* **8**, 195–210.

Efron, B. and R. Tibshirani (1984). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–77.

Helmholtz, H. F. (1885/1954). *On the Sensations of Tone.* New York: Dover.

Joos, M. (1948). Acoustic phonetics. In L. S. of America (Ed.), *Language Monograph No 23.* Baltimore, MD: Waverly.

Lindblom, B. E. F. (1963, nov). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America* **35**(11), 1773–1781.

Majid, R., C. Abry, L.-J. Boe, and P. Perrier (1987). Fonctions de macro-sensibilité articulatori-acoustique, pour le système vocalique du français, à partir d'un modèle articulatoire du conduit vocal. *Bulletin du Laboratoire de la Communication Parlée de Grenoble* **1A**, 133–157.

Perrier, P., C. Abry, and E. Keller (1990). Vers une modélisation des mouvements du dos de la langue. *Journal d'Acoustique* **2**, 69–77.

Pitermann, M. and J. Caelen (1993, september). Comparison of the variability of formants and

formant targets using dynamic modelling. In *Proceedings of Eurospeech '93*, Volume 1, Berlin, Germany, pp. 397–400. catalyst consult: ESCA.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1988). *Numerical Recipes in C*, Chapter 14: Modeling of Data, pp. 540–547. Cambridge University Press.

Strange, W. (1987). Information for vowels on formant transitions. *Journal of Memory and Language* **26**, 550–557.

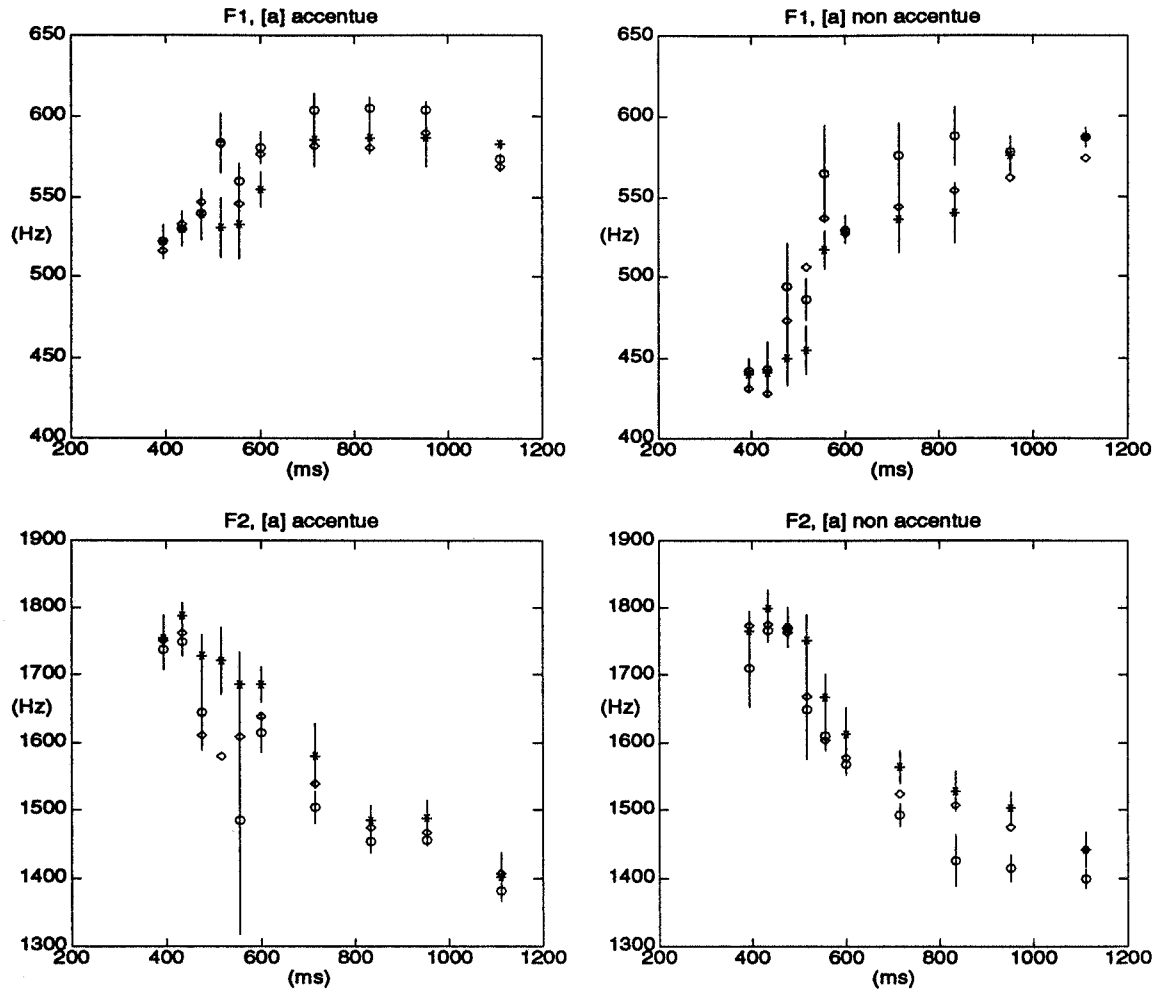


FIG. 2 - *Fréquences formantiques (*), estimations de cible par modèle linéaire (o) et par modèle sigmoïdal (◇) en fonction de l'intervalle de temps séparant deux impulsions du métronome.*

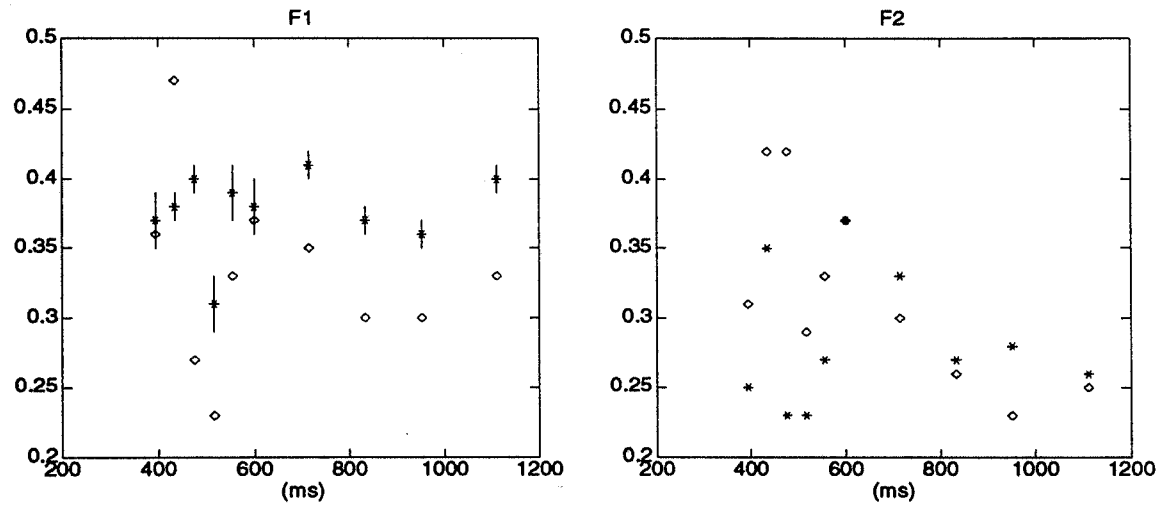


FIG. 3 - *Coefficient de relaxation de la sigmoïde pour le [a] accentué (*) et non accentué (◇) en fonction de l'intervalle de temps séparant deux impulsions du métronome.*

PROTECTION DE SEGMENTS SUB-PHONETIQUES EN SYNTHESE PSOLA

PIERRE-YVES LE MEUR

Elan Informatique
4, rue Jean Rodier Z.I. de Montaudran 31400 Toulouse
CNET LAA/TSS/RCP
2, route de Trégastel 22301 Lannion Cedex

Résumé

L'algorithme TD-PSOLA réduit la durée d'un signal de parole en supprimant des blocs de signal. Appliqué "brutalement" en synthèse de parole par concaténation d'unités préenregistrées, pour la réduction de la durée d'un phonème, ce traitement peut provoquer la suppression partielle ou totale d'événements sub-phonémiques. Cet article présente une méthode de protection de segments sub-phonétiques lors du traitement PSOLA de la durée. L'application de cette méthode pour la protection des occlusives sourdes, en synthèse du français, est décrite en détail. Elle se décompose en deux étapes: après marquage automatique du début de la barre d'explosion, seule la durée du silence est modifiée en synthèse. L'efficacité de cette méthode de protection des occlusives sourdes du français est enfin évaluée par un test subjectif d'écoute.

Mots-clés

Synthèse PSOLA, Traitement de durée, Evénements sub-phonémiques.

1. INTRODUCTION

Le système de synthèse de parole du CNET est basé sur la concaténation d'unités préenregistrées. Il utilise l'algorithme TD-PSOLA pour modifier, par phonème, la durée et la fréquence fondamentale des unités. En appliquant une déformation uniforme tout au long du phonème, le traitement de la durée peut entraîner la disparition ou la dégradation d'événements sub-phonémiques cruciaux pour l'intelligibilité. C'est pourquoi nous proposons une méthode de protection de certains segments sub-phonétiques de toute déformation en durée.

La première partie de cet article rappelle le principe de l'algorithme TD-PSOLA et montre l'intérêt de protéger des segments sub-phonétiques lors du traitement de la durée. Dans une deuxième partie, une application à la protection des occlusives sourdes du français est décrite. Enfin, la dernière partie présente les résultats d'un test subjectif d'écoute.

2. PROTECTION DE SEGMENTS SUB-PHONETIQUES LORS DU TRAITEMENT PSOLA DE LA DUREE

2.1. PRINCIPE DE L'ALGORITHME TD-PSOLA

L'algorithme TD-PSOLA (Time Domain Pitch-Synchronous Overlap-Add) [1] est une technique de synthèse qui permet la modification, dans le domaine temporel, de la durée et de la fréquence fondamentale d'un signal de parole. Son principe est la superposition et l'addition de blocs de signal décalés dans le temps.

Dans une première étape, l'algorithme transforme le signal original x(n) en une suite de signaux élémentaires d'analyse $x_m(n)$, par multiplication du signal original x(n) par une suite de fenêtres d'analyse $h_m(n)$ de type Hanning:

$$x_m(n) = h_m(t_m - n)\, x(n)$$

La longueur des fenêtres d'analyse $h_m(n)$ est telle que deux signaux élémentaires consécutifs présentent un recouvrement mutuel important (typiquement 50 %). Les instants t_m sont appelés marques d'analyse. Sur les portions voisées du signal, elles se succèdent à cadence synchrone de la fréquence fondamentale. Sur les portions non voisées, elles se succèdent à intervalle de temps régulier (typiquement 10 ms).

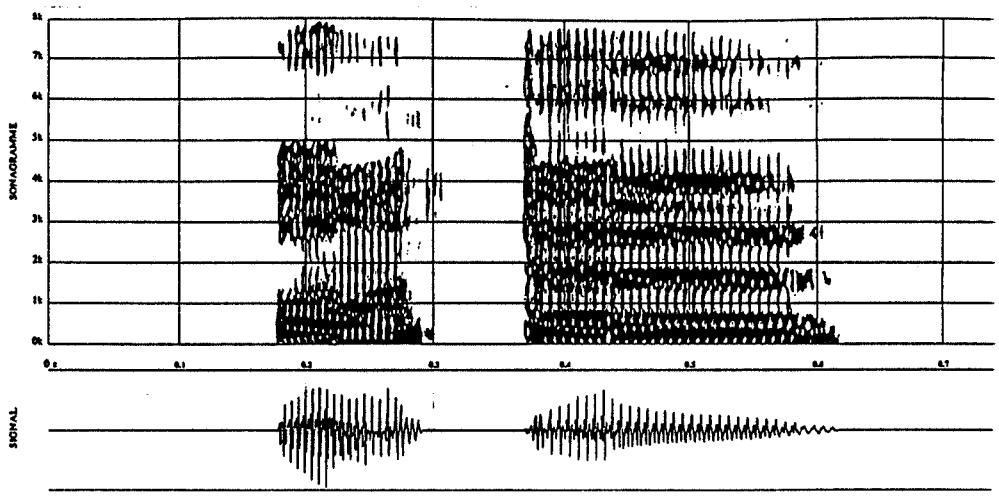


Figure 1
Exemple de suppression partielle de la tenue de l'occlusion et de la barre d'explosion de l'occlusive /t/ sur le signal de synthèse du mot /õtø/

Dans une deuxième étape, une suite de signaux élémentaires de synthèse $x'_q(n)$ est produite à partir de la suite des signaux élémentaires d'analyse $x_m(n)$. La suite des signaux élémentaires $x'_q(n)$ est synchronisée sur une nouvelle suite d'instants t'_q, appelés marques de synthèse. Chaque signal élémentaire de synthèse $x'_q(n)$ provient d'un signal élémentaire d'analyse unique $x_m(n)$, déterminé par une fonction de déformation temporelle $t'_q \rightarrow t_m$ reliant les marques de synthèse aux marques d'analyse:

$$x'_q(n) = x_m(n + t_m - t'_q)$$

La dernière étape consiste à calculer le signal de synthèse x'(n) par simple superposition et addition des signaux élémentaires de synthèse $x'_q(n)$:

$$x'(n) = \Sigma\, x'_q(n)$$

Le signal de synthèse x'(n) apparaît donc comme une somme de versions pondérées et translatées du signal original x(n).

2.2. PROTECTION DE SEGMENTS SUB-PHONETIQUES

En définissant des délais à appliquer entre les signaux élémentaires d'analyse, la fonction de déformation temporelle $t'_q \rightarrow t_m$ permet la modification de la fréquence fondamentale du signal original. Par suppression ou duplication de signaux élémentaires d'analyse, cette fonction permet la modification de la durée du signal original. C'est à ce mécanisme de modification de la durée du signal original, et plus particulièrement à la réduction de sa durée, que nous allons nous intéresser.

Dans le système de synthèse de parole du CNET, la déformation de la durée d'un phonème est uniforme: la sélection des signaux élémentaires à supprimer dans le cas d'une accélération du phonème (ou à dupliquer dans le cas d'un ralentissement du phonème) est répartie uniformément sur toute la durée du phonème. Des tests d'intelligibilité de la parole synthétisée par le système du CNET [2] ont fait apparaître un pourcentage non négligeable d'erreurs sur les occlusives, plus particulièrement sur les occlusives sourdes.

Une explication réside dans les modalités de mise en oeuvre du traitement PSOLA de la durée. En effet, la durée intrinsèque du phonème X (telle que stockée dans le répertoire de diphones) est la somme de la durée du demi-phonème X dans le diphone AX et de la durée du demi-phonème X dans le diphone XB. Suivant le contexte lexical et syntaxique, la durée intrinsèque de ce phonème sera ensuite réduite (ou allongée) dans des proportions qui peuvent être importantes. Par exemple, si un phonème doit être accéléré d'un facteur deux, l'algorithme PSOLA supprimera un signal élémentaire sur deux, tout au long de ce phonème. Une conséquence néfaste de ce traitement "brutal" est la suppression partielle ou totale d'événements sub-phonémiques, telle que la tenue de l'occlusion ou la barre d'explosion des occlusives.

La figure 1 présente un tel exemple. La durée intrinsèque du phonème /t/ est de 171 ms (96 ms dans le

diphone /õt/ et 75 ms dans le diphone /tø/), alors que la durée requise du phonème /t/ lors de la synthèse du mot /õtø/ ("honteux") est de 111 ms: l'application "brutale" de PSOLA pour obtenir une telle réduction de durée entraîne la suppression partielle de la tenue de l'occlusion et de la barre d'explosion du phonème /t/.

C'est pourquoi nous proposons de marquer, sur les unités préenregistrées, des segments sub-phonétiques à protéger lors du traitement PSOLA de la durée. Le paragraphe suivant présente une application de ce marquage à la protection des occlusives sourdes du français.

3. APPLICATION A LA PROTECTION DES OCCLUSIVES SOURDES

3.1. PRINCIPE DE LA METHODE

Une occlusive se compose d'une suite d'événements acoustiques: un silence qui correspond à la phase d'occlusion du conduit vocal, une barre d'explosion qui correspond au relâchement de l'air et un bruit de friction. Pour les occlusives sourdes du français (/p/, /t/, /k/), nous proposons une nouvelle fonction de réduction de la durée, qui tient compte des événements acoustiques.

La méthode de protection mise au point se décompose en deux étapes. Tout d'abord, un algorithme marque automatiquement le début de la barre d'explosion des occlusives sur les diphones XB, la lettre X désignant l'une des trois occlusives sourdes et la lettre B un phonème quelconque. La marque de début d'explosion est posée entre la marque début (marque située dans le silence de l'occlusive X) et la marque milieu (marque séparant l'occlusive X du phonème B) du diphone XB. Cette marque est maintenue sur le diphone XB du répertoire. A la synthèse, quand les diphones AX et XB sont concaténés, la marque de début d'explosion sépare l'occlusive X en deux segments sub-phonétiques distincts: le premier correspond au silence (le demi-phonème X dans le diphone AX et la première partie du demi-phonème X dans le diphone XB) et le second à la barre d'explosion et au bruit de friction (la seconde partie du demi-phonème X dans le diphone XB). L'algorithme PSOLA modifie de façon non uniforme la durée des occlusives sourdes, en distinguant les deux segments séparés par la marque de début d'explosion.

3.2. AUTOMATISATION DU MARQUAGE SUB-PHONETIQUE

Le marquage sub-phonétique d'une occlusive sourde est réalisé automatiquement. La détection du début de l'explosion s'appuie sur une analyse acoustique du signal des logatomes. Un vecteur acoustique est calculé toutes les 8 ms et sur une trame de durée 16 ms: ses coefficients résultent d'une analyse en banc de filtres sur une échelle MEL [3].

Pour chaque trame de l'occlusive, on déduit de son vecteur acoustique associé, l'énergie dans les hautes fréquences E_{hf}, la barre d'explosion se caractérisant par une énergie importante dans les hautes fréquences. Partant de la trame dont l'énergie E_{hf} est maximale, on progresse de droite à gauche jusqu'à la trame dont l'énergie E_{hf} est inférieure à 20 % de l'énergie E_{hf} maximale. La marque de début d'explosion est posée au centre de la trame suivante.

3.3. MODIFICATION NON UNIFORME DE LA DUREE

Lors de l'étape de modification de la fréquence fondamentale d'un phonème par l'algorithme PSOLA, un coefficient de déformation en fréquence α_m est associé à chaque marque d'analyse t_m. Ce coefficient correspond à un facteur de déformation du fondamental. Lors de l'étape de modification de la durée, un coefficient de déformation en durée β_m est associé à chaque marque d'analyse t_m. Ce coefficient est constant pour toutes les marques d'analyse du phonème et correspond au rapport entre la durée cible et la durée intrinsèque du phonème. Le signal élémentaire, correspondant à la marque d'analyse t_m, n'est ni dupliqué ni supprimé si $\alpha_m \beta_m = 1$.

Dans [4], un coefficient d'élasticité est associé à chaque signal élémentaire, de façon à dupliquer ou supprimer en priorité les signaux élémentaires des segments stables: ce coefficient est pris proportionnel à la fonction d'interpolation, tronquée à 50 %, des cibles d'une décomposition temporelle [5].

Pour protéger un segment sub-phonétique de toute déformation en durée, on prend $\beta_m = 1 / \alpha_m$ sur ce segment. La différence entre la durée cible et la durée effectivement réalisée par le segment protégé, est alors répartie uniformément sur l'ensemble des segments non protégés. Dans le cas des phonèmes non voisés, $\alpha_m = 1$. On protège donc un segment d'une occlusive sourde en prenant $\beta_m = 1$ sur ce segment.

Après marquage sub-phonétique des occlusives sourdes, le second segment constitué de la barre d'explosion et du bruit de friction est alors protégé de toute réduction en durée lors du traitement PSOLA. Sur le premier segment, les cinq derniers signaux élémentaires sont également protégés. Ainsi, il n'est pas possible de réduire la tenue de l'occlusion à moins de 50 ms.

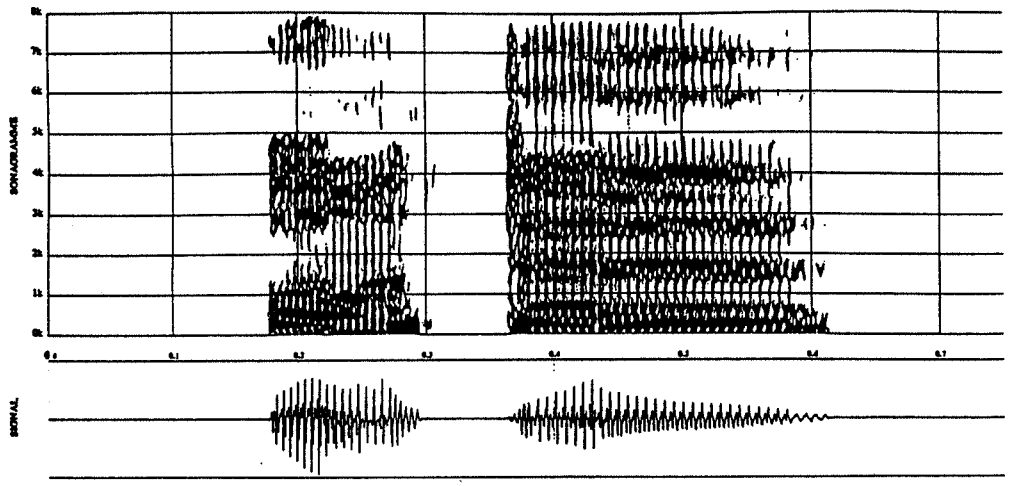


Figure 2
Exemple de protection de l'occlusive /t/ sur le signal de synthèse du mot /õtø/

La figure 2 présente l'exemple de la protection de l'occlusive /t/ sur le signal de synthèse du mot /õtø/. La durée intrinsèque du phonème /t/, 171 ms, est réduite à 111 ms, durée spécifiée par les traitements "hauts-niveaux". Sur cette figure, la barre d'explosion, protégée de toute réduction en durée, est plus marquée que sur la figure 1, qui correspond à la synthèse "sans protection" du même mot.

4. EVALUATION DE LA PROTECTION DES OCCLUSIVES SOURDES

4.1. PROTOCOLE D'EVALUATION

Pour évaluer l'apport de la protection des occlusives sourdes sur la qualité de la parole synthétique, nous avons effectué un test subjectif d'écoute. Le type de test retenu est la comparaison par paires ("test ABBA") [6]: les paires de stimuli, présentées dans l'ordre AB et BA, correspondent à une même expression synthétisée avec (version A) et sans (version B) la protection des occlusives sourdes. L'auditeur exprime sa préférence pour la première ou la seconde présentation du stimulus, ou l'équivalence des deux présentations.

Dix auditeurs français non naïfs se sont soumis à trois séances de test, chaque séance ayant pour objectif d'évaluer l'apport de la protection des occlusives sourdes sur un répertoire de diphones particulier. Les répertoires correspondent à deux voix d'homme (voix RG et OB) et une voix de femme (voix FE) en français. Les trois répertoires ont été segmentés automatiquement [7].

Les stimuli, synthétisés par le système CNETVOX-PSOLA du CNET [8], sont différents d'une séance à l'autre, donc d'un répertoire à l'autre. Pour l'ensemble du test, 65 stimuli ont été présentés aléatoirement dans l'ordre AB et dans l'ordre BA: 21 stimuli synthétisés avec la voix RG, 20 stimuli avec la voix OB et 24 stimuli avec la voix FE. Il s'agit de mots isolés ou de courtes expressions contenant au moins une occlusive sourde.

Ils ont été choisis en fonction du contexte et de la position de l'occlusive dans l'expression. Quatre groupes sont à distinguer:
- le premier groupe (groupe VCV) est constitué de 18 mots (pour les 3 répertoires, soit 6 mots en moyenne par répertoire), 7 mots contenant une occlusive sourde en position initiale et suivie d'une voyelle (exemple: /tymœR/), 5 mots contenant une occlusive sourde en position finale et précédée d'une voyelle (exemple: /*e*f*ɛ*k/), et 6 mots contenant une occlusive sourde intra-mot en contexte intervocalique (exemple: /matine/),
- le deuxième groupe (groupe VCCV1) est constitué de 20 expressions (pour les 3 répertoires, soit 7 expressions en moyenne par répertoire) contenant un groupe consonantique en contexte intervocalique, composé de deux occlusives sourdes apparaissant de part et d'autre d'une frontière de mots (exemple: /yn nap tise/),
- le troisième groupe (groupe VCCV2) est constitué de 15 mots (pour les 3 répertoires, soit 5 mots en moyenne par répertoire) contenant un groupe consonantique en contexte intervocalique, composé d'une occlusive sourde et d'une consonne d'une autre classe (exemple: /RøpRiz/),

- le dernier groupe (groupe VCCCV) est constitué de 12 mots (pour les 3 répertoires, soit 4 mots en moyenne par répertoire) contenant un groupe consonantique en contexte intervocalique, composé d'au moins trois consonnes dont au moins une occlusive sourde (exemple: /ɛkspeRiãs/).

Enfin, il faut remarquer que pour la moitié des stimuli (33 stimuli), la protection de l'occlusive sourde entraîne le non-respect de la consigne prosodique de durée émanant des traitements "hauts-niveaux": la durée effectivement réalisée à la synthèse dépasse de plus de 10 ms (un signal élémentaire) la durée spécifiée par le module prosodique de CNETVOX. Pour les voix RG et OB (mêmes consignes prosodiques), l'allongement de chaque occlusive sourde pour ces 33 stimuli est en moyenne de 27,1 ms; pour la voix FE, il est de 28,5 ms. Une section du paragraphe suivant présente les résultats du test en distinguant les stimuli qui respectent la consigne de durée (groupe R) de ceux qui ne la respectent pas (groupe NR).

4.2. RESULTATS ET ANALYSE

4.2.1. Résultat global

Les 10 auditeurs ont écouté 130 paires de stimuli. Le tableau 1 présente le résultat global du test.

TYPE DE REPONSE	SCORE (%)
Sans protection	4,6
Avec protection	12,3
Equivalent	34,5
Incohérent	48,6

Tableau 1

Sont considérées comme "incohérentes", les réponses où le choix d'un auditeur varie avec l'ordre de présentation d'un stimulus dans une paire. Malgré des scores "Equivalent" et "Incohérent" importants dus à la difficulté du test, l'écart des scores "Avec protection" vs "Sans protection" (7,7 %) montre une préférence pour la protection des occlusives sourdes.

4.2.2. Résultats par type de voix

Les 10 auditeurs ont écouté 42 paires de stimuli synthétisées avec la voix RG, 40 paires avec la voix OB et 48 paires avec la voix FE. Le tableau 2 présente le score obtenu par chaque groupe de stimuli.

	SCORE (%)		
TYPE DE REPONSE	RG	OB	FE
Sans protection	3,8	4,0	5,8
Avec protection	17,1	9,0	10,8
Equivalent	24,8	41,0	37,5
Incohérent	54,3	46,0	45,9

Tableau 2

Les écarts des scores "Avec protection" vs "Sans protection" (RG: 13,3 %; OB: 5,0 %; FE: 5,0 %) montrent tous une préférence pour la protection des occlusives sourdes. De plus, on remarque que les scores obtenus par chaque type de réponse sont du même ordre de grandeur: la tendance des réponses ne semble donc pas dépendre du répertoire de diphones avec lequel les stimuli sont synthétisés.

4.2.3. Résultats en fonction du contexte

Les 10 auditeurs ont écouté 36 paires de stimuli du groupe VCV, 40 paires du groupe VCCV1, 30 paires du groupe VCCV2 et 24 paires du groupe VCCCV. Le tableau 3 présente le score obtenu par chaque groupe de stimuli.

	SCORE (%)			
TYPE DE REPONSE	VCV	VCCV1	VCCV2	VCCCV
Sans protection	5,6	4,5	5,3	2,5
Avec protection	7,8	20,0	12,0	6,7
Equivalent	37,2	25,5	36,7	42,5
Incohérent	49,4	50,0	46,0	48,3

Tableau 3

Pour les groupes consonantiques, les écarts des scores "Avec protection" vs "Sans protection" (VCCV1: 15,5 %; VCCV2: 6,7 %; VCCCV: 4,2 %) montrent une préférence pour la protection des occlusives sourdes. Pour le groupe VCCV1 (groupe consonantique composé de deux occlusives sourdes apparaissant de part et d'autre d'une frontière de mots), l'importance de l'écart peut s'expliquer par la surarticulation généralement observée en naturel entre les mots, surarticulation qui est mieux respectée à la synthèse "avec protection" grâce au respect de la durée minimale (50 ms) imposée au silence de la seconde occlusive sourde. Dans les cas VCCV2 et VCCCV, la présence d'une liquide au sein du groupe consonantique a pu gêner la comparaison des stimuli, du fait de sa qualité médiocre en synthèse.

4.2.4. Résultats en fonction du critère de respect de la consigne de durée

Les 10 auditeurs ont écouté 64 paires de stimuli du groupe R (consigne de durée respectée) et 66 paires du groupe NR (durée allongée de 28 ms en moyenne par rapport à la consigne de CNETVOX). Le tableau 4 présente le score obtenu par chaque groupe de stimuli.

	SCORE (%)	
TYPE DE REPONSE	R	NR
Sans protection	4,0	5,1
Avec protection	8,1	16,4
Equivalent	39,1	30,0
Incohérent	48,8	48,5

Tableau 4

Pour chaque groupe, l'écart des scores "Avec protection" vs "Sans protection" (R: 4,1 %; NR: 11,3 %) montre une préférence pour la protection des occlusives sourdes. Les résultats observés sur les stimuli de type NR ne permettent pas de préciser si c'est l'allongement global de la durée des occlusives ou la protection de zones qui est responsable de l'amélioration perçue par les auditeurs. Par contre, les résultats obtenus sur les stimuli de type R sont eux directement attribuables à la protection de zones, puisque les durées des occlusives sont rigoureusement identiques dans les deux cas, "avec" et "sans protection". On peut en outre remarquer que, pour les stimuli du groupe R, les durées des occlusives (identiques à celles prédites par CNETVOX) sont en moyenne plus longues (21 ms) que celles des occlusives apparaissant dans les stimuli du groupe NR (pourtant allongées par rapport à celles prédites par CNETVOX). Puisque l'effet de la "protection" reste sensible dans ces cas (stimuli de type R), on peut faire l'hypothèse que, même en cas d'allongement global de la durée des occlusives, la protection de zones est bénéfique et a contribué à l'amélioration perçue par les auditeurs.

5. CONCLUSION

L'algorithme PSOLA réalise la réduction de la durée d'un son par la suppression de signaux élémentaires. Si cette opération est répartie uniformément sur toute la durée d'un phonème, elle peut entraîner la disparition ou la dégradation d'événements sub-phonémiques. Nous avons proposé une stratégie de déformation qui permet, après marquage de certains événements sub-phonémiques, de les protéger de toute déformation en durée lors du processus de synthèse. Une application à la protection des occlusives sourdes du français est présentée: après marquage du début de la barre d'explosion, celle-ci est protégée de toute déformation en durée, et une durée minimale est imposée à la tenue de l'occlusion. L'évaluation par un test subjectif de cette méthode montre que la protection des occlusives sourdes apporte une certaine amélioration. Une autre conclusion de ce test est que les durées des occlusives sourdes telles qu'actuellement prédites par le module prosodique CNETVOX mériteraient sans doute d'être allongées dans certains contextes (en particulier les groupes consonantiques). Cette méthode de protection de segments sub-phonétiques pourrait être étendue aux autres classes de phonèmes: par exemple, il pourrait être intéressant de protéger la transition formantique entre voyelles acoustiquement distantes.

Références

[1]: Moulines E., Charpentier F., 1990,
"Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones",
Speech Communication, vol. 9, pp. 453-467

[2]: Emerard F., Cartier M., 1991,
"Test d'intelligibilité de systèmes de synthèse à partir du texte",
Note technique CNET, NT/LAA/TSS/427

[3]: Zwicker E., Feldtkeller R., 1981,
"Psychoacoustique: l'oreille, récepteur d'information",
Collection technique et scientifique des Télécommunications, Masson

[4]: Bailly G., et al., 1992,
"Automatic labelling of large prosodic databases: tools, methodology, and links with a text-to-speech system"
Talking Machines, pp. 323-333

[5]: Bimbot F., 1988,
"Synthèse de la parole: des segments aux règles, avec utilisation de la décomposition temporelle",
Thèse de l'ENST

[6]: Calliope, 1989,
"La parole et son traitement automatique",
Collection technique et scientifique des Télécommunications, Masson

[7]: Boeffard O., 1993,
"Segmentation automatique d'unités acoustiques pour la synthèse de la parole",
Thèse de l'Université de Rennes I

[8]: Moulines E., et al., 1990,
"A real-time French Text-to-Speech system generating high-quality synthetic speech",
Proc. ICASSP, Albuquerque, pp. 309-312

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.

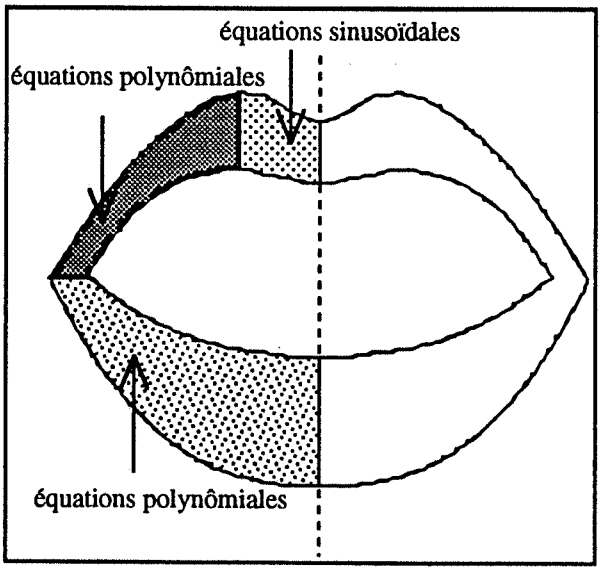


Figure 1 : Les trois parties du modèle de lèvres 2D (Guiard-Marigny, 1992).

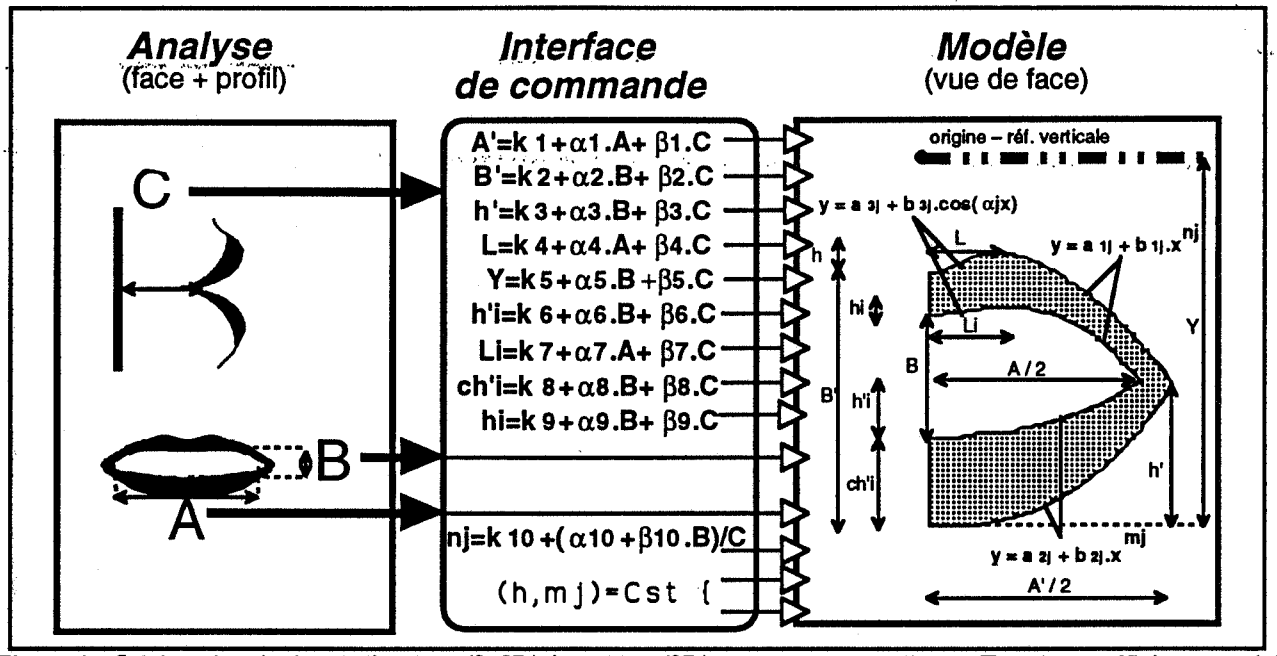


Figure 2 : Schéma de principe de l'analyse (2x2D) / synthèse (2D) des contours des lèvres. Tous les coefficients ont été établis expérimentalement de façon à ajuster au mieux les contours modélisés avec les contours réels (Guiard- Marigny, 1992).

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.

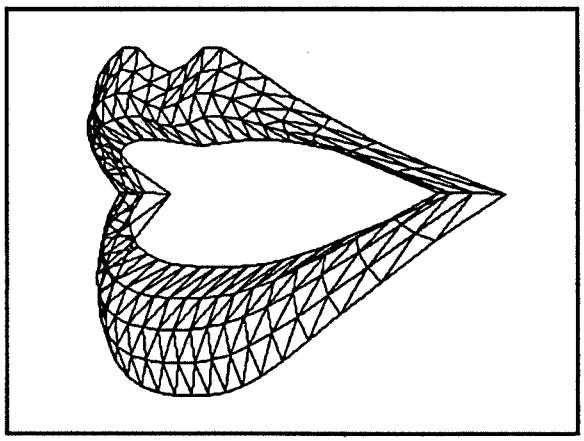


Figure 3 : Représentation en fil de fer du modèle 3D - vue de trois-quarts (Adjoudani, 1993).

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).

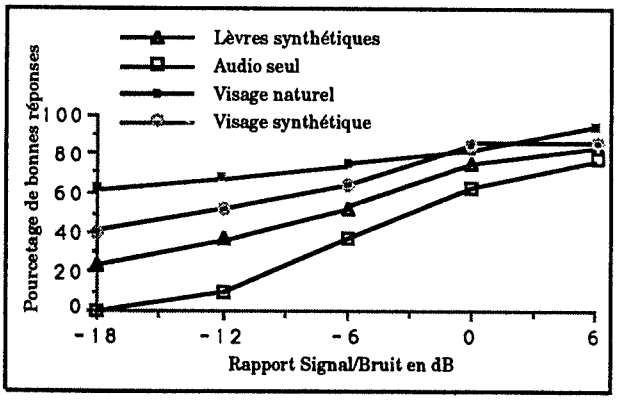


Figure 4 : Courbes représentant les résultats de l'expérience de Le Goff (1993) sur l'apport de l'information visuelle à l'intelligibilité de la parole

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/, /J/, /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).

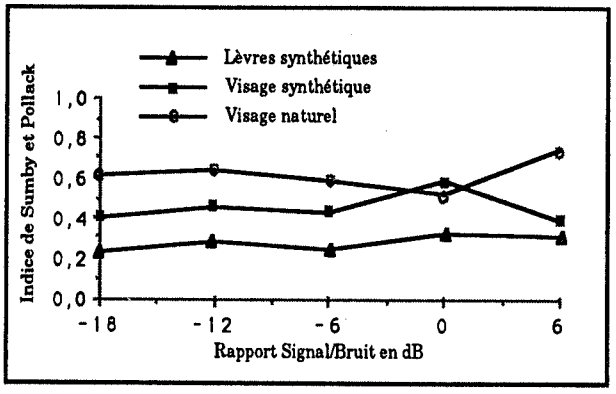


Figure 5 : Proportions de l'intelligibilité acoustique manquante compensée par l'information visuelle (indice défini pour la première fois par Sumby et Pollack (1954) : (I[AV]-I[A])/(1-I[A]).

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 a 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.

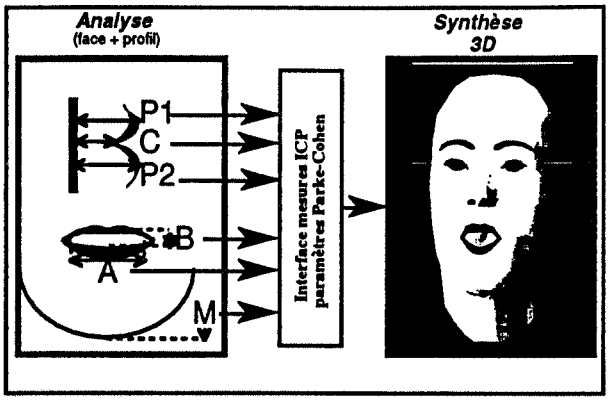


Figure 1 : Principe de l'analyse/synthèse du visage paramétrique (Le Goff, 1993).

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).

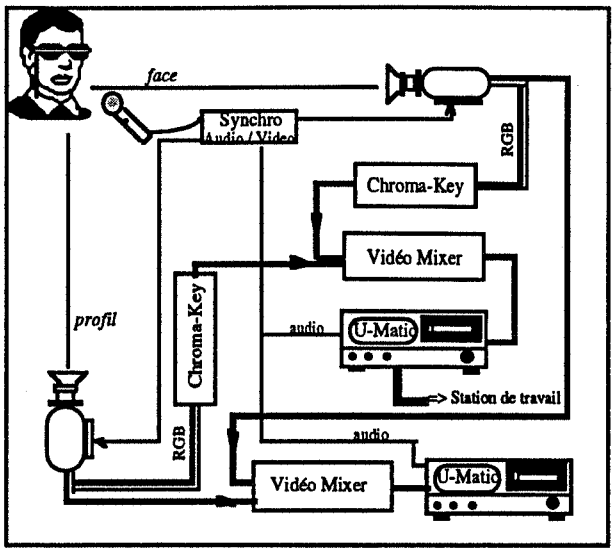


Figure 2 : Le poste "vidéo-parole" (Lallouache, 1991).

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.

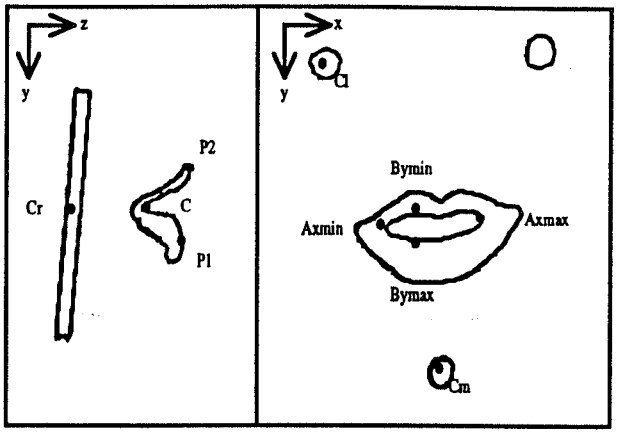


Figure 4 : Contours des objets détectés sur une image vidéo pour le calcul des paramètres à transmettre (Angola, 1993).

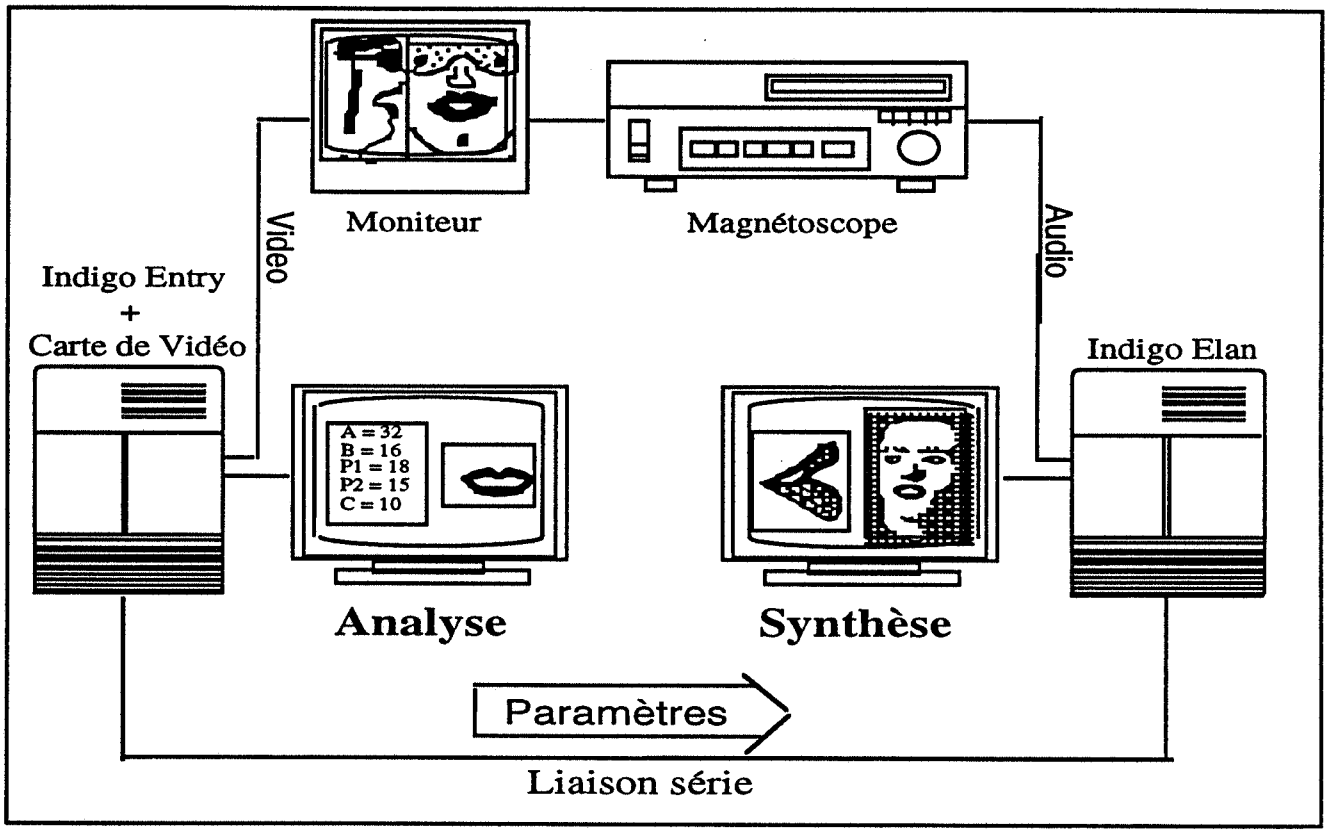


Figure 5 : Schéma de principe du système d'analyse/synthèse en temps réel d'un visage parlant (Angola, 1993).

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.

SESSION AFFICHEE : A2

SEGMENTATION ET RECONNAISSANCE

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

UTILISATION D'UN RESEAU NEURO-FLOU POUR LA DISTINCTION DU TRAIT VOISE/NON-VOISE

ROBERT SOKOL (1), PIERRE-YVES GLORENNEC (2)
GUY MERCIER (3), KRZYSZTOF WOLINSKI (4)

(1) FRANCE TELECOM, CNET LANNION, LAA/TSS/SMD
ROUTE DE TREGASTEL
22301 LANNION, FRANCE

(2) INSA
DEPARTEMENT D'INFORMATIQUE
35000 RENNES, FRANCE

(3) FRANCE TELECOM, CNET LANNION, LAA/TSS/RCP
ROUTE DE TREGASTEL
22301 LANNION, FRANCE

(4) UNIVERSITE DE RENNES I / IFSIC
35000 RENNES, FRANCE

Résumé: Dans cette communication nous décrivons une application des réseaux neuro-flous à la détection du trait voisé/non-voisé sur un signal de parole de qualité téléphonique.
Dans un premier temps nous faisons un court rappel sur la logique floue ainsi que sur la possibilité d'inclure des éléments de cette logique dans un réseau neuronal.
Nous parlons ensuite du réseau obtenu et nous présentons l'algorithme d'apprentissage utilisé, ainsi que les résultats de reconnaissance sur le trait voisé/non-voisé que nous utilisons pour séparer deux groupes de consonnes (b, d, g, v, z, m, n) et (p, t, k, f, s).

Mots clés: réseaux neuronaux, logique floue, traits phonétiques, reconnaissance de la parole

1. INTRODUCTION

Dans les travaux que nous menons nous avons voulu évaluer l'apport des méthodes neuro-floues dans un système de reconnaissance de la parole. Pour celà nous avons intégré des règles issues d'un système à base de connaissances, KEAL [1], dans une structure de type neuronal.
Cette structure fait partie des réseaux neuro-flous et plus particulièrement d'un réseau développé par P.Y. Glorennec et K. Wolinski à l'IRISA [2].
Ce couplage réalisé, nous avons utilisé le réseau obtenu afin de distinguer le trait voisé/non-voisé sur un signal de parole de qualité téléphonique.

2. SYSTEME D'INFERENCE FLOUE

2.1 LOGIQUE FLOUE

Les bases de cette logique ont été introduites par Zadeh [3].

Par la suite, de nombreux travaux ont été réalisés et des applications concrètes ont vu le jour.

La notion de base de la logique floue est la notion de degré d'appartenance à un ensemble. Nous pouvons diviser le domaine de variation d'une variable $u, u \in U$, en sous ensembles, A, caractérisés par une fonction d'appartenance $\mu_A : u \rightarrow \mu_A(u) \in [0,1]$.

La valeur $\mu_A(u)$ est le degré d'appartenance de u à l'ensemble A. La notion de **degré d'appartenance** permet de manipuler des notions imprécises comme "grand", "petit", "éloigné" etc.

Un système d'inférence floue à n entrées et une sortie est caractérisé par:

- une **base de connaissance** exprimée sous forme de règles du type

règle i: **si** x_1 est A_{i1} ... **et** x_n est A_{in} **alors** y est B_i

où les x_i sont les variables d'entrée, y est la variable de sortie, les A_{ij} et B_i, des sous-ensembles flous.

- une méthode d'évaluation des entrées par rapport aux sous-ensembles flous (**fuzzification**), et une méthode de passage du flou au réel pour la sortie y (**défuzzification**),

- un **mécanisme d'inférence** permettant d'évaluer l'ensemble des règles et d'en tirer une conclusion.

Dans le cas d'un système à deux entrées, où les sous-ensembles B_i sont des valeurs exactes (non floues), notées w_i, la prise de décision est faite à l'issue des étapes suivantes, cf [4].

1- Pour une entrée donnée (x_1, x_2), on calcule les différents degrés d'appartenance aux sous-ensembles flous apparaissant dans la partie prémisse des règles. Sur la figure 1 nous avons montré un exemple de fonctions d'appartenance pour la variable x_1 sur son domaine de variation X_1. Les degrés d'appartenance aux ensembles A_1 et A_2 sont les images de x_1 par des fonctions triangulaires.

2- Pour chaque règle, on calcule **la valeur de vérité**, α_i, telle que:

$$\alpha_i = ET\{\mu_{A_{i1}}(x_1), \mu_{A_{i2}}(x_2)\}$$

où l'opérateur ET est un opérateur de conjonction en logique multi-valuée.

Cette valeur de vérité est aussi appelée le degré d'appariement de l'observation avec la règle i.

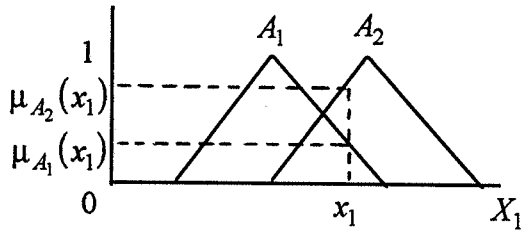


figure 1: fonctions A_1 et A_2
degrés d'appartenance $\mu_{A_1}(x_1)$ et $\mu_{A_2}(x_1)$

3- Pour la deffuzification nous allons utiliser la méthode du centre de gravité: les w_i des parties conclusion des règles sont pondérés par les valeurs de vérité relatives des prémisses:

$$y = \frac{\sum \alpha_i w_i}{\sum \alpha_i}$$

Ce système peut être représenté par une structure de type connexionniste (figure 2). Grâce à cette représentation nous remarquons une structure en couches, chacune de ces couches ayant une fonction précise.

La couche 1 redirige les entrées vers les unités de la couche 2. La couche 2 calcule les degrés d'appartenance aux sous-ensembles flous. La couche 3 calcule les valeurs de vérité des règles et, enfin, l'unité de la couche 4 calcule la sortie y.

Ce système d'inférence floue est un approximateur universel [5]

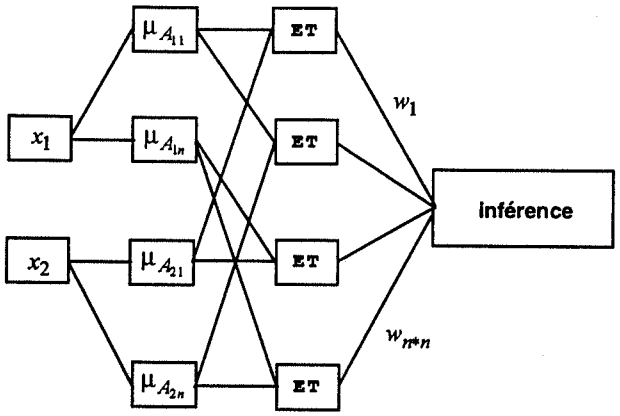


figure 2:
structure connexioniste d'un système d'inférence floue

2.2 LE RESEAU NEURO-FLOU UTILISE

Nous allons particulariser le système précédent pour nous rapprocher d'un réseau neuronal ordinaire. Pour cela, nous allons faire les choix suivants :

- Les fonctions d'appartenance sont gaussiennes. Leur implémentation neuronale est précisée en annexe.
- L'opérateur de conjonction, ET, est une version lissée de la conjonction de Lukasiewicz.

Soit ET(A,B) la valeur de vérité de la prémisse "x est A ET y est B". La conjonction de Lukasiewicz s'écrit :

$$ET(A,B) = Max(0, \mu_A(x) + \mu_B(y) - 1)$$

Cette expression peut s'exprimer par un neurone dont le vecteur de poids est (1,1,-1), avec la fonction d'activation u -> Max(0,u). Cependant, cette fonction n'étant pas dérivable en zéro, nous allons l'approximer par la sigmoide d'équation

$$u \rightarrow S(u) = \frac{1}{1 + \exp\left(-\frac{x - 0.5}{b}\right)}$$

Le coefficient b est choisi tel que $y(0)=1-y(1)=\varepsilon$. En prenant $\varepsilon = 0.056$, la fonction u -> S(u) est telle que :

$$\sup_{u \in [-1,1]} |Max(0,u) - S(u)| = 0.056$$

On prend donc:

$$b = (2Log((1-\varepsilon)/\varepsilon)^{-1} \approx 0.177$$

Dans ce cas, la valeur de vérité d'une règle est implémentée par un neurone "ordinaire", de vecteur de

poids (1/b,1/b,-1.5/b), et dont la fonction d'activation est une sigmoide.

2.3 ALGORITHME D'APPRENTISSAGE

Etant donné que le neurone de sortie dans le réseau utilisé calcule le **barycentre** de ses entrées (defuzzification), il faut modifier l'algorithme de rétropropagation classique.

Nous donnons ces modifications pour un réseau de type Lukasiewicz.

Soit $N(n,i)$ le $i^{ème}$ neurone de la couche n. La formule de modification des poids s'écrit usuellement:

$$\Delta w_{ij}^n = -\eta \delta_i^n o_j^{n-1}$$

avec:

w_{ij}^n = poids synaptique entre $N(n,i)$ et $N(n-1,j)$

δ_i^n = terme d'erreur de $N(n,i)$

o_j^{n-1} = sortie de $N(n-1,j)$

a_j^n = activation de $N(n,j)$

$y_i = i^{ème}$ sortie du réseau et d_i la valeur désirée correspondante.

Les termes d'erreur pour les couches 2, 3 et 4 sont alors donnés par les formules suivantes:

$$\delta_i^4 = \frac{-d_i + y_i}{\sum_j o_j^3}$$

$$\delta_i^3 = f'(a_i^3) \sum_k \delta_k^4 (w_{ki}^4 - y_k)$$

$$\delta_i^2 = g'(a_i^2) \sum_k \delta_k^3 w_{ki}^3$$

f et g étant les fonctions d'activation des couches 3 et 2.

3. DU FLOU AU NEURONAL ET RECIPROQUEMENT

Par le choix de la conjonction de Lukasiewicz, le système d'inférence floue représenté en figure 2 est devenu un réseau neuronal ordinaire, avec des connexions incomplètes. Nous allons faire subir à ce système deux opérations successives.

3.1 DU FLOU AU NEURONAL

En établissant une interconnexion totale entre les couches 2 et 3, nous transformons le système flou initial en un réseau neuronal à accès semi-distribué [6]. En ce faisant, nous perdons la sémantique des règles associée aux neurones de la couche 3. Cet établissement de connexions peut se faire soit en initialisant les poids suppl.émentaires à zéro et en les laissant évoluer au cours de l'apprentissage, [2], soit en les initialisant aléatoirement.

3.2 DU NEURONAL AU FLOU

Pour retrouver les règles ainsi que l'interprétation des poids 2→3 et 3→4 nous pouvons procéder à un élagage automatique des poids ayant peu d'influence sur le taux d'erreur. Cette influence peut être évaluée grâce au rapport:

$$\frac{E(w_{ij} \neq 0) - E(w_{ij} = 0)}{E(w_{ij} \neq 0)}$$

$E(...)$ étant l'erreur de classification.

Ce rapport était calculé pour tous les poids 2→3. Nous supprimons ensuite tous les poids pour lesquels ce rapport est inférieur ou égal à zero.

Ce passage dans les deux sens, flou-neuronal, a été évoqué de manière très vague dans [7]. Son interêt est :

- d'une part, d'initialiser un réseau neuronal en y introduisant une éventuelle connaissance a priori. Cela facilite donc l'apprentissage.
- d'autre part, d'extraire la connaissance acquise sous la forme de r.ègles floues.

4. APPLICATION AU TRAITEMENT DE LA PAROLE

4.1 PRESENTATION DE KEAL

KEAL est un système de reconnaissance de la parole continue dépendant du locuteur. Il possède une structure hiérarchique qui combine des éléments statistiques, structurels et des connaissances sous forme de règles. Par exemple le décodage acoustico-phonétique est basé sur des règles qui segmentent le signal de parole en phonèmes et fournit les principales caractéristiques phonétiques de ces segments. En parallèle des fonctions discriminantes dépendantes du locuteur permettent d'ajuster ces premières hypothèses.

4.2 REGLES ET VECTEUR D'ENTREE

Dans KEAL, les règles concernent les relations entre les énergies des différents canaux d'un vocodeur à 14 canaux. Toutes ces règles sont du type:

règle i: **si** x_1 est A_{i1} ... **et** $x_n$est A_{in} **alors** y est B_i

où les x_i sont les variables composant le vecteur d'entrée. Dans le cas du trait voisé/non-voisé ces règles s'appliquent toutes au niveau de la trame. Les variables composant le vecteur d'entrée sont au nombre de 9 et sont constituées par différentes combinaisons des énergies des 14 canaux. Soit $K'_{f_1-f_2}$ l'énergie dans

l'intervalle de fréquence $f_1 - f_2$ à l'instant t. Les 9 variables sont:

$$K^t_{250-450}$$

$$K^t_{250-450} - K^t_{650-850}$$

$$K^t_{250-450} - K^t_{450-650}$$

$$\max_{850-1050 \le \Delta f \le 3900-4300} K^t_{\Delta f} - K^t_{250-450}$$

$$\max_{850-1050 \le \Delta f \le 3900-4300} K^t_{\Delta f} - K^t_{450-650}$$

$$K^t_{250-450} - K^t_{450-650} - K^t_{650-850}$$

$$K^t_{250-450} - K^{t+1}_{250-450}$$

$$\sum_{\Delta f=3900-4300} K^t_{\Delta f} - \sum_{\Delta f=1050-1300} K^t_{\Delta f}$$

$$\sum_{\Delta f=3100-3500} K^t_{\Delta f} - \sum_{\Delta f=250-450} K^t_{\Delta f}$$

$$2 * K^t_{250-450} - K^t_{450-650} - K^t_{650-850}$$

Nous disposons d'un ensemble de 15 règles pour le trait voisé/non-voisé.

4.3 INITIALISATION DES POIDS

Nous pouvons initialiser les poids entre la couche d'entrée et la première couche cachée grâce aux règles issues de KEAL. Etant donné que nous utilisons des gaussiennnes nous initialisons leurs moyennes avec les seuils utilisés dans les règles.

Nous avons vu que le vecteur d'entrée est constitué de 9 composantes. Ces composantes sont normalisées sur l'intervalle [-1, 1]. Etant donné que pour les composantes 4, 5, 6, 7, 8 nous n'avons qu'un ensemble flou par variable, nous avons ajouté un ensemble aux domaines de variation de ces composantes. Nous obtenons donc 29 valeurs de degrés d'appartenance. Ce nombre est égal au nombre de neurones de la première couche cachée.

Les degrès d'appartenance à ces ensembles sont calculés par des fonctions gaussiennes dont les moyennes ont été initialisées grâce aux valeurs fournies par les règles de KEAL.

Pour gagner en robustesse et en propriété d'interpolation nous avons établi une interconnexion totale entre les neurones de la première couche cachée et ceux de la deuxième. En établissant cette interconnexion nous perdons la signification physique des poids 2→3 et 3→4. Cependant nous pourrons récupérer leur signification en élagant, une fois l'apprentissage terminé, les poids 2→3 ayant peu d'importance sur le taux d'erreur. Le nombre de cellules de la deuxième couche cachée (égal au nombre de règles) s'élève à 15.

Les poids partant de cette couche vers le neurone de sortie sont initialisés à 0 (nous supposons que toutes les règles sont d'égale importance).

4.4 APPRENTISSAGE

Les essais ont été effectués sur la base du Trégor qui est constituée par un vocabulaire de 36 mots d'un serveur vocal interactif de renseignement sur la région de Lannion. Ces mots ont été prononcés par 760 locuteurs. Les frontières des phonèmes et leurs étiquettes sont fournis par un Modèle de Markov [8]. Nous avons divisé cette base en trois ensembles. Le premier constitue **l'ensemble d'apprentissage** , le deuxième **l'ensemble de cross-validation** et le dernier **l'ensemble de test**. On entraîne le réseau à l'aide de l'ensemble d'apprentissage, et regulièrement on teste sa faculté de généralisation sur l'ensemble de cross-validation. Nous arrêtons l'apprentissage dès que le taux d'erreur sur cet ensemble augmente. On utilise alors l'ensemble de test afin de vérifier le pouvoir généralisateur du réseau obtenu.

L'algorithme d'apprentissage a été décrit en 5. La sortie du réseau est forcée à 0 en présence de trames voisées (phonèmes b, d, g, v, z, m, n) et à 1 en présence de trames sourdes (p, t, k, f, s).

5. RESULTATS

Les tableaux 2, 3, 4 donnent les résultats obtenus au niveau de la trame ainsi qu'au niveau du segment, la segmentation étant fournie par un Modèle de Markov [8]. Les résultats ont été obtenus en modifiant sélectivement les poids entre les différentes couches.

Ainsi pour obtenir les résultats du tableau 2 nous modifions uniquement les poids w_i (poids 3→4) de la partie conclusion des règles. Les résultats du tableau 3 sont obtenus avec une modification des coefficients w_i et des paramètres des fonctions d'appartenance (poids 1→2, 3→4). Quant au tableau 4 nous modifions l'ensemble des poids (poids 1→2, 2→3, 3→4). Au cours de ces essais le pas d'apprentissage η est constant et égal à 0,05. Les poids entre les couches 2 et 3 sont initialisés à -1 ou +1, le signe étant généré de manière pseudo-aléatoire.

La décision quant à l'affectation de la trame à l'ensemble des trames voisées ou non-voisées est prise grâce à un seuil de décision. Au niveau du segment la décision est prise en calculant la moyenne des sorties des trames comprises entre le début et la fin du segment. Cette moyenne est ensuite comparée au seuil de décision.

Pour retrouver les règles nous procédons à l'élagage des poids 2→3. Après un premier élagage nous avons pu diminuer de moitié le nombre de liaisons 2→3 sans diminuer les performance sur l'ensemble de test. En continuant l'élagage nous avons obtenu un réseau neuro-flou avec 36 règles 21 fonctions d'appartenance. Les performances obtenues sont donnés dans le tableau 5.

6. PERSPECTIVES

Nous pouvons également appliquer cette méthode de travail aux coefficients du cepstre qui présentent l'avantage d'être décorrelés ce qui n'est pas le cas lors de l'utilisation des énergies.
Nous voulons également étendre nos essais à d'autres traits tels que voyelle/consonne et utiliser d'autres méthodes d'élagage moins coûteuses en temps de calcul comme l'Optimal Brain Damage [9]. Ce travail est actuellement en cours.

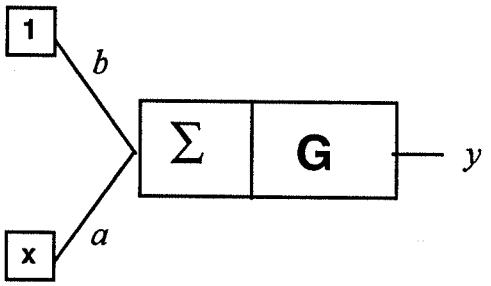
7. CONCLUSION

Les résultats obtenus sont du même ordre de grandeur que ceux obtenus sur le même ensemble de phonèmes avec un T.D.N.N. (Time Delay Neural Network) [10] 10,5 à 10,0% de taux d'erreur au niveau du segment en fonction de la paramètrisation du réseau pendant l'apprentissage [11].
L'avantage du réseau neuro-flou réside dans l'interprétabilité de ses paramètres et la possibilité d'inclure des connaissances a priori dans les poids et l'architecture.

8. ANNEXE

8.1 NEURONE GAUSSIEN

Une fonction gaussienne, $x \rightarrow \exp(-x^2)$, est très simple à implémenter. Soit ε la valeur résiduelle de la courbe de Gauss aux points $x = m \pm dx$. La donnée du triplet (m ,b, ε) permet donc de déterminer immédiatement les coefficients a et b :



$$y = e^{-(ax+b)^2};\ a = \frac{\sqrt{Log\frac{1}{\varepsilon}}}{dx};\ b = -am$$

figure 5: neurone gaussien

8.2 LES REGLES DE KEAL

Dans KEAL, les règles concernent les relations entre les énergies des différents canaux d'un vocodeur à 14 canaux. Toutes ces règles sont du type:

règle i: **si** x_1 est A_{i1} ... **et** $x_n$est A_{in} **alors** y est B_i

En voici 3 exemples:

$K'_{250-450} < 4dB \Rightarrow$ trame non-voisée

$K'_{250-450} - K'_{650-850} < 0dB \Rightarrow$ trame non-voisée

$K'_{250-450} \geq 8dB \cap K'_{250-450} - K'_{450-650} \geq 0dB \Rightarrow$ trame voisée

9. REFERENCES

[1] G. Mercier et Al., "Recognition of Speaker Dependent Continuous Speech with KEAL", Readings in Speech Recognition, Edited by A. Waibel et K.F. Lee, Morgan Kaufmann Publishers, 1990.

[2] P.Y. Glorennec, "Un réseau neuro-flou évolutif", Actes de Neuro-Nîmes'91, novembre 91.

[3] L.A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes", IEEE Trans. on MSC, vol 3, 1973.

[4] M. Sugeno, "Fuzzy control", Nikkan Kogyo, Tokyo, 1988.

[5] P. Y. Glorennec, "A general class of Fuzzy Inference Systems: Application to identification and control", 2th European System Science Congress, Prague, october 1993.

[6] V. Lorquet, P. Puget, JJ. Niez, "Réseaux multicouches à accès semi-distribués", Neuro-Nîmes, 1991.

[7] R. Masuoka, N. Watanebe, A. Kawamura, Y. Owada, K. Asakawa, "Neuro-fuzzy system", Proc. of Iisuka91, Iisuka, Japon 1991.

[8] D. Jouvet, "Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques", Thèse E.N.S.T, 1988.

[9] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal Brain Damage", Advances in Neural Information Processing (2), D. S. Touretzky, Ed. (Denver 1989) 1990.

[10] A. Waibel et Al., "Phoneme Recognition Using Time Delay Neural Networks", Proc. IEEE, Int. Conf. ASSP, 1989.

[11] B. Paille, "Identification de traits phonétiques à l'aide de réseaux connexionnistes", Rapport de stage N° 751 LAA/TSS/RCP, juillet 1993.

	trame		segment	
	sourde	voisée	sourd	voisé
erreur	18,9%		15,1%	

tableau 2: modification des poids 3→4

	trame		segment	
	sourde	voisée	sourd	voisé
erreur	14,9%		10,3%	

tableau 3: modification des poids 1→2, 3→4

	trame		segment	
	sourde	voisée	sourd	voisé
erreur	14,2%		9,9%	

tableau 4: modification des poids 1→2, 2→3, 3→4

	trame		segment	
	sourde	voisée	sourd	voisé
erreur	14,3%		10,0%	

tableau 5: réseau neuro-flou, modification des poids 1→2, 3→4

COMPARAISON DES MULTIMODELES ET DES DENSITES MULTIGAUSSIENNES POUR LA RECONNAISSANCE DE LA PAROLE PAR MODELES DE MARKOV

D. JOUVET, M. DAUTREMONT, A. GOSSART

FRANCE TELECOM, CNET LANNION, LAA/TSS/RCP
ROUTE DE TREGASTEL, 22300 LANNION, FRANCE

Résumé

L'augmentation du nombre de paramètres des modèles de Markov conduit généralement à une meilleure modélisation du signal de parole, qui se traduit par de meilleures performances en reconnaissance. L'accroissement du nombre de paramètres peut s'effectuer dans plusieurs directions : utilisation de plusieurs modèles pour chaque unité à reconnaître ou emploi de densités multigaussiennes. Ces deux approches sont comparées sur plusieurs corpus. L'apport des multimodèles est présenté et analysé. Les résultats montrent que les techniques d'augmentation progressive du nombre de paramètres des modèles sont très efficaces, tant pour l'accroissement du nombre de modèles des multimodèles que pour l'augmentation du nombre de gaussiennes des densités multigaussiennes.

Mots clés : *Modèles de Markov, multimodèles, densités multigaussiennes, classification des locuteurs, apprentissage incrémental.*

INTRODUCTION

Dans le contexte de la modélisation markovienne du signal de parole, l'augmentation du nombre de paramètres des modèles conduit généralement à une meilleure modélisation du signal de parole, et par conséquence à de meilleures performances en reconnaissance. Pour les modèles de Markov à densités gaussiennes, l'accroissement du nombre de paramètres peut s'effectuer dans deux directions : soit en augmentant le nombre de modèles représentant chaque unité à reconnaître, soit en augmentant le nombre de gaussiennes, ce qui se traduit généralement par l'emploi de densités multigaussiennes.

L'emploi de plusieurs modèles pour chaque unité à reconnaître conduit à une modélisation plus précise. Chaque modèle est appris à partir d'un sous-ensemble des données d'apprentissage. Le choix de ces sous-ensembles caractérise les différentes techniques de multimodélisation. Les données correspondant à chaque mot peuvent être classifiées en comparant les formes acoustiques [1] ou bien en appliquant des techniques similaires à la quantification vectorielle, c'est-à-dire en effectuant des divisions et perturbations de modèles déjà appris [2]. Après chaque modification une nouvelle classification des données est effectuée, et les classes restent figées pour la suite de l'apprentissage. Une autre possibilité consiste à initialiser différemment les modèles à partir de variantes de prononciation (par exemple normale, forte, lente, ...), puis à laisser l'apprentissage affecter individuellement chaque observation au modèle le plus "proche" [3]. Dans la littérature, on trouve fréquemment l'emploi de bimodèles avec une classification a priori des locuteurs entre hommes et femmes. Toutefois, dans certaines conditions, la classification par accent peut être plus pertinente [4].

Puisqu'un modèle affecté à un groupe restreint et homogène de locuteurs est normalement plus discriminant qu'un modèle appris sur l'ensemble des locuteurs, une structure hiérarchique de modèles, de plus en plus spécifiques au fur et à mesure que l'on s'éloigne de la racine de l'arbre à été étudiée [5]. Cette approche permet de débuter la reconnaissance avec le modèle le plus général, ensuite le locuteur est affecté au groupe le plus "adéquat" et la reconnaissance est poursuivie avec le modèle correspondant à ce groupe. De cette approche, nous avons retenu la méthode de classification des locuteurs (en phase d'apprentissage) qui prend en compte l'ensemble des données disponibles pour chaque locuteur. Pendant la reconnaissance, les suites de modèles affectés aux différentes classes sont mises en parallèle.

L'emploi de densités de probabilité d'émission multigaussiennes évite toute hypothèse a priori sur la forme de la densité, pourvu que le nombre de gaussiennes soit suffisamment élevé. L'apprentissage des gaussiennes de chaque densité est généralement effectué par une technique reposant sur la classification explicite des trames affectées à chaque densité en autant de classes qu'il y a de gaussiennes à estimer [6]. Toutefois le nombre de composantes de chaque densité peut être ajusté de manière dynamique [7] en fonction des données d'apprentissage. De manière plus générale, l'adaptation dynamique de la structure des modèles de Markov permet, soit de réduire le nombre de gaussiennes par regroupements successifs afin de rendre le modèle exploitable en temps réel, soit d'accroître de manière sélective, par "division et perturbation", le nombre de gaussiennes afin d'améliorer les

performances en reconnaissance [8]. C'est cette dernière approche qui a été reprise pour la mise en oeuvre des multigaussiennes. Le seul critère imposé lors de la "division" d'une gaussienne porte sur le nombre de trames qui lui ont été affectées lors de l'apprentissage ; ce nombre doit être supérieur à un seuil pré-défini.

Après une présentation du système de reconnaissance et de la modélisation contextuelle par allophones, nous décrivons les corpus de données qui ont servi pour l'évaluation et la comparaison des divers algorithmes. Ensuite, nous présentons les multimodèles et les variantes d'apprentissage disponibles, en insistant sur la technique de classification automatique des locuteurs. Puis, nous décrivons les densités multigaussiennes et la technique d'apprentissage choisie, et nous en présentons l'apport au niveau des performances de reconnaissance. Finalement, nous comparons l'apport respectif des multimodèles et des densités multigaussiennes en fonction de leur taille exprimée par le nombre total de gaussiennes.

SYSTEME DE RECONNAISSANCE

Le système de reconnaissance de la parole employé dans cette étude repose sur une modélisation markovienne du signal de parole. La version de base de ce système, Phil90, fut développée essentiellement pour satisfaire les applications vocales dans le domaine des télécommunications [9], et en particulier la commande vocale des serveurs vocaux interactifs.

L'analyseur acoustique calcule toutes les 16 ms un ensemble de 8 coefficients cepstraux ainsi que le logarithme de l'énergie de la trame. Les coefficients cepstraux sont obtenus à partir d'une transformée en cosinus du logarithme des énergies dans 24 filtres répartis selon une échelle Mel et couvrant les fréquences inférieures à 4 kHz. A cet ensemble de coefficients acoustiques, on ajoute une approximation des dérivées temporelles premières et secondes. En conséquence la modélisation markovienne traite des vecteurs de 27 coefficients [10]. Dans Phil90, les fonctions de densité de probabilité d'émission des observations sont représentées par des gaussiennes avec des matrices de covariance diagonales, et sont liées aux transitions de la chaîne de Markov.

Phil90 utilise un réseau entièrement compilé qui inclut les modèles de silence de début et de fin de phrase et autorise divers types de modélisations. Les unités peuvent être des mots, des phonèmes, des allophones, ... Le compilateur reçoit en entrée toutes les connaissances a priori sur la tâche : la syntaxe qui précise les suites de mots autorisées, la description des mots en phonèmes (ou autres unités choisies), les règles phonologiques à appliquer sur ces descriptions, et enfin la structure des modèles acoustiques. A partir de toutes ces informations, le compilateur génère un réseau dans lequel chaque chemin de l'état initial à l'état final correspond à une phrase valide de la tâche traitée.

L'algorithme de Viterbi est appliqué en phase de reconnaissance pour trouver le meilleur chemin dans ce réseau. En phase d'apprentissage, l'algorithme de Viterbi est également employé, mais il n'est appliqué qu'à la portion du réseau qui représente les différentes variantes possibles de prononciation de la phrase traitée.

Pour la reconnaissance de petits vocabulaires en mode mots isolés, la modélisation directe de chaque mot (sans le décomposer en unités plus petites) fournit de très bonnes performances. Pour le traitement de mots connectés, l'utilisation d'unités phonétiques avec une modélisation contextuelle pertinente devient préférable. C'est cette modélisation par allophones [11] qui sera employée dans la plupart des expériences présentées dans cet article. Elle consiste à partager de nombreux paramètres entre les modèles des diverses réalisations d'un phonème, en particulier la modélisation des "cibles" est commune. Ainsi avec un nombre restreint de paramètres on obtient une bonne modélisation contextuelle des divers allophones d'un son donné.

CORPUS DE PAROLE

Tous les corpus de données de parole utilisés ont été acquis à travers le réseau téléphonique, puis validés par écoute. Chaque "enregistrement" contient la parole prononcée par le locuteur (mot ou phrase), plus un peu de silence avant et après le mot.

Les corpus des chiffres (de 0 à 9), du Trégor (36 mots ou expressions employés dans des serveurs vocaux) et des nombres (de 00 à 99) ont été enregistrés par près de 800 locuteurs répartis dans différentes villes de France (essentiellement des communications téléphoniques interurbaines). Ils couvrent donc divers accents régionaux. Le corpus des 500 mots (les plus fréquents du français) a été enregistré par 10 locuteurs. Chaque locuteur ayant prononcé trois fois la totalité des 500 mots au cours de plusieurs appels téléphoniques.

Tableau 1 - Nom et taille des corpus de parole

Corpus de parole (et nombre de locuteurs)	Nombre d'enregistrements	
	Apprentissage	Test
Chiffres (770 locuteurs)	3 555	3 622
Trégor (765 locuteurs)	12 720	12 844
Nombres (789 locuteurs)	7 304	7 288
500 mots (10 locuteurs)	9 766	4 883

Chaque corpus est découpé en deux parties, l'une pour l'apprentissage des paramètres des modèles et l'autre pour le test. En ce qui concerne le corpus des 500 mots, 2 répétitions de chaque locuteur sont utilisées pour l'apprentissage, la troisième est réservée pour le test et fournit des performances de reconnaissance en mode plurilocuteur. Pour les trois autres corpus (chiffres, Trégor et nombres), les locuteurs de l'ensemble de test sont différents de ceux de l'ensemble d'apprentissage, il s'agit donc de performances en mode

indépendant du locuteur. Le tableau 1 ci-dessus, résume les tailles des corpus de parole, et leur découpage.

Pour la plupart des expériences décrites dans cet article, les coefficients acoustiques employés sont les coefficients cepstraux bruts résultant de l'analyse acoustique "standard". Toutefois, pour les expériences de classification automatique des locuteurs, nous avons voulu nous affranchir autant que possible de la dispersion due aux lignes téléphoniques. Dans ce but, nous avons effectué ces expériences de classification des locuteurs avec des données compensées, c'est-à-dire après adaptation aux lignes téléphoniques. La technique d'adaptation employée consiste à soustraire le cepstre à long terme [12] estimé sur l'ensemble des données provenant d'un même appel et à normaliser le coefficient énergie. Cette technique d'adaptation est plus efficace que les techniques de filtrage passe-haut, mais elle ne se prête pas à un fonctionnement temps-réel.

STRUCTURES MULTIMODELES

Dans le cas d'une multimodélisation, chaque mot du vocabulaire est représenté par plusieurs modèles. L'option choisie consiste à dupliquer autant de fois que nécessaire l'ensemble du réseau. Cette précision prend toute son importance dans le cas de la reconnaissance de phrases, ou lorsque l'on emploie des unités de reconnaissance de taille inférieure au mot (allophones par exemple).

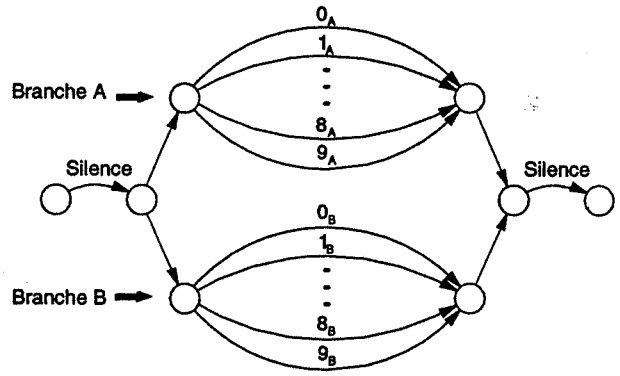


Figure 1 - Structure d'un bimodèle pour les chiffres.

La figure ci-dessus schématise un bimodèle dans le cas des chiffres. Chacune des branches (A et B) correspond à une copie du réseau et dispose de son propre jeu de paramètres. Les modèles des silences (de début et de fin de phrase) sont communs aux deux branches. 0_A et 0_B représentent les deux modèles associés au chiffre 0. En phase de reconnaissance, l'algorithme de Viterbi recherchera le meilleur chemin dans ce réseau, chacune des branches étant également probable. Par contre en phase d'apprentissage, diverses variantes sont possibles suivant l'affectation des données de parole aux branches du multimodèle.

APPRENTISSAGE STANDARD

Cette variante est la plus simple. Lors de l'apprentissage, chaque observation (résultant de la prononciation d'un mot ou d'une phrase) peut être affectée à l'une ou l'autre des branches indépendamment des autres données. De fait, elle sera automatiquement affectée à la branche qui fournit la plus forte probabilité d'émission. C'est l'apprentissage standard des modèles de Markov lorsque plusieurs prononciations d'un mot sont possibles, l'algorithme de Viterbi sélectionne la plus probable à chaque itération de l'apprentissage.

CLASSIFICATION A PRIORI DES LOCUTEURS

Cette variante consiste à associer chaque branche à un ensemble de locuteurs défini de manière arbitraire. Une classification courante consiste à séparer les données en deux classes, l'une correspondant aux données des hommes et l'autre à celles des femmes. Les données des hommes serviront à l'apprentissage des paramètres des modèles de la branche A par exemple, et les données des femmes à l'apprentissage des paramètres de la branche B. Cette partition des données est conservée pendant toute la durée de l'apprentissage.

Cette classification a priori en deux classes conduit à de bonnes performances de reconnaissance comme on le verra plus loin, et elle est fréquemment employée. L'inconvénient de la partition arbitraire des données en deux classes réside dans la difficulté d'étendre l'approche de manière simple et efficace à un nombre de classes supérieur à deux.

CLASSIFICATION AUTOMATIQUE DES LOCUTEURS

Une classification automatique des locuteurs supprime cette restriction. On peut créer autant de classes que l'on désire. L'algorithme d'apprentissage développé repose sur une procédure itérative. On crée autant de classes de locuteurs qu'il y a de branches dans le multimodèle. Chaque locuteur sera alors affecté automatiquement à l'une des classes. L'ensemble des données de la classe servira ensuite à estimer les paramètres des modèles de la branche correspondante. Puis on répétera ces deux étapes jusqu'à la convergence.

L'idée principale de cette méthode consiste à ne pas séparer les données provenant d'un même locuteur, ce qui permet de modéliser une "certaine corrélation" entre ces données. Plusieurs méthodes sont disponibles pour affecter un locuteur à une classe. S'agissant de données d'apprentissage, donc connues, on peut utiliser soit un critère de vraisemblance soit un critère de taux d'erreur. C'est ce dernier qui a été retenu. Chaque locuteur est affecté à la branche conduisant au plus faible taux d'erreur (en cas d'égalité, on utilise la branche la plus probable). Cette affectation favorise le critère "taux d'erreur" ce que ne fait pas l'apprentissage standard.

ANALYSE DES PERFORMANCES DE RECONNAISSANCE

La figure 2 ci-dessous présente les performances de reconnaissance des bimodèles sur différents corpus.

L'axe vertical indique les taux d'erreur. L'emploi d'une échelle logarithmique facilite les comparaisons. Avec le modèle standard par allophones, les taux d'erreur et les intervalles de confiance à 95 % sont 1,13 % [± 0,34] pour les chiffres, 0,84 % [± 0,15] pour le Trégor, 3,46 % [± 0,42] pour les nombres à 2 chiffres et 6,72 % [± 0,70] pour les 500 mots.

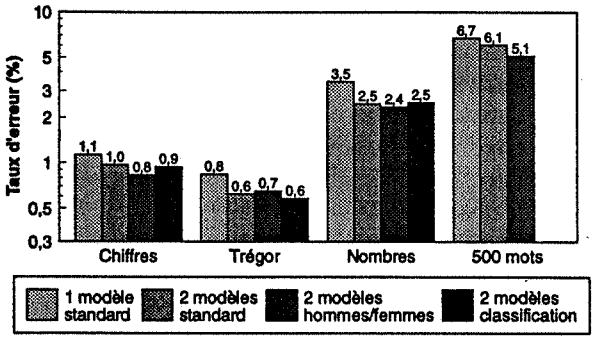


Figure 2 - Comparaison des diverses modélisations sur les données brutes (i.e. sans adaptation aux lignes).

Quel que soit le corpus, les bimodèles conduisent à de meilleures performances de reconnaissance. Le taux d'erreur est réduit de 10 à 30 % suivant les corpus et la technique d'apprentissage employée. Tantôt c'est la classification a priori hommes/femmes qui conduit aux meilleures performances, tantôt c'est la classification automatique. Au vu des résultats il est difficile de conclure à un net avantage d'une technique d'apprentissage par rapport aux autres. Elles sont sensiblement équivalentes.

Afin de mieux cerner l'impact des techniques d'apprentissage, la même expérience a été également menée en employant les données compensées, c'est-à-dire après adaptation aux lignes téléphoniques.

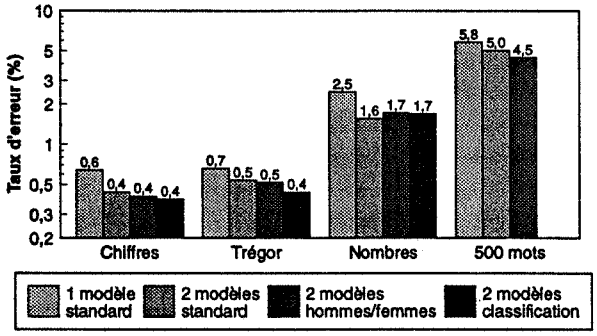


Figure 3 - Comparaison des diverses modélisations sur les données compensées (i.e. adaptation aux lignes).

Là encore tous les bimodèles fournissent de meilleures performances que le modèle de base, et on ne peut pas conclure à une nette suprématie d'une technique d'apprentissage par rapport aux autres. On note toutefois un résultat intéressant concernant la méthode par classification automatique. Elle conduit aux mêmes performances que la classification a priori hommes/femmes sur les corpus des chiffres et des nombres, et fournit de meilleures performances sur le corpus du Trégor.

AU-DELA DE 2 CLASSES

On peut donc envisager d'étendre la technique de classification automatique des locuteurs pour apprendre des multimodèles avec plus de 2 branches. La figure ci-dessous présente les taux d'erreur obtenus sur le corpus des chiffres, en utilisant les données compensées, pour deux types de modélisations (mots et allophones) et des étapes différentes lors de l'apprentissage : soit on apprend directement le nombre voulu de modèles, soit on augmente progressivement le nombre des modèles.

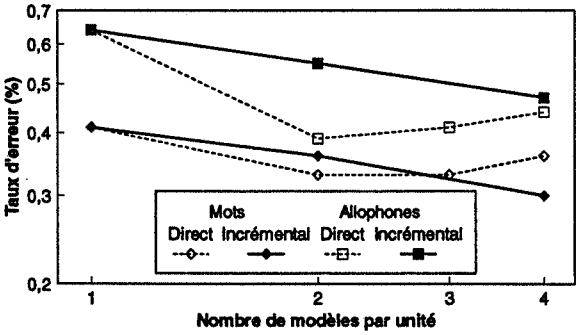


Figure 4 - Comparaison des méthodes d'apprentissage des multimodèles sur le corpus des chiffres avec les données compensées (i.e. adaptation aux lignes).

Les courbes en tirets donnent les taux d'erreur obtenus lorsque le nombre de classes voulues est fixé a priori et que toutes les classes sont estimées simultanément. Les courbes en trait plein suivent une approche incrémentale. On apprend d'abord correctement un monomodèle. A partir de ce monomodèle on initialise les paramètres des branches du bimodèle (valeurs des paramètres du monomodèle, plus légère perturbation aléatoire), puis on effectue l'apprentissage du bimodèle. Puis on répète le processus si nécessaire. A partir du bimodèle on initialise le quadrimodèle (chaque branche du bimodèle servant à initialiser deux branches du quadrimodèle) puis on effectue l'apprentissage (toujours par classification automatique des locuteurs).

L'augmentation progressive du nombre de modèles conduit à un comportement beaucoup plus régulier. Chaque étape de division (puis d'apprentissage) entraîne une réduction de 12 à 15 % du taux d'erreur.

Pour le quadrimodèle par allophones obtenu par la méthode incrémentale, 2 classes contiennent une majorité de locuteurs hommes (respectivement 79 % et 90 %), et les 2 autres une majorité de locuteurs femmes (respectivement 64 % et 72 %). On note également que la partition de l'ensemble d'apprentissage est équilibrée, il y a sensiblement autant de locuteurs dans chaque classe (respectivement 93, 100, 84 et 105).

DENSITES MULTIGAUSSIENNES

Une approche fréquemment employée pour raffiner la modélisation markovienne consiste à employer des densités de probabilité multigaussiennes. Avec un nombre suffisant de gaussiennes, on peut approcher n'importe quelle distribution.

METHODE D'APPRENTISSAGE

Considérant, pour des raisons de stabilité des procédures, qu'il vaut mieux d'abord apprendre un modèle avec peu de paramètres avant de raffiner la modélisation, la procédure mise en oeuvre pour l'apprentissage des densités multigaussiennes repose sur une augmentation progressive du nombre de gaussiennes de chaque densité. Elle constitue en fait une extension de la modification dynamique des structures des modèles [13]. Ainsi au lieu de diviser quelques gaussiennes seulement, on divise toutes les gaussiennes de la modélisation, puis on perturbe légèrement les nouveaux paramètres avant de reprendre l'apprentissage. Le nombre théorique de gaussiennes est multiplié par deux à chaque étape. Des contraintes sur le nombre minimal de trames acoustiques affectées à chaque gaussienne (quelques dizaines) peuvent conduire à un écart entre ce nombre théorique et le nombre moyen effectif de trames par densité. Cet écart est d'autant plus faible que le corpus d'apprentissage est grand.

PERFORMANCES DE RECONNAISSANCE

La figure suivante indique les taux d'erreurs obtenus sur les corpus des nombres et des 500 mots (données brutes) en fonction du nombre moyen de gaussiennes par densité. Les courbes en tirets indiquent les performances sur les ensembles d'apprentissage, et les courbes en trait plein les performances sur les ensembles de test. Afin de ne pas alourdir la figure on n'a pas représenté les résultats sur les corpus des chiffres et du Trégor, le comportement étant similaire.

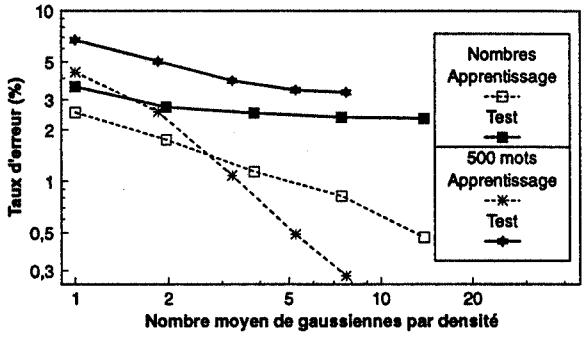


Figure 5 - Influence du nombre moyen de gaussiennes par densité (performances sur données brutes).

La première conclusion qui s'impose est l'amélioration des performances de reconnaissance en fonction de l'augmentation du nombre de gaussiennes par densité. A signaler toutefois que l'amélioration est beaucoup plus importante sur les données d'apprentissage que sur les données de test. L'écart de performances entre l'apprentissage et le test est d'autant plus important que le nombre de paramètres des modèles est élevé. Pour un nombre théorique de 16 gaussiennes par densité, la réduction du taux d'erreur sur les données d'apprentissage est supérieur à 80 % pour les 4 corpus traités. En ce qui concerne les ensembles de test, la réduction du taux d'erreur est de 35 % pour les corpus des chiffres, du Trégor et des nombres, et de 50 % pour le corpus des 500 mots.

La taille du corpus d'apprentissage des 500 mots est insuffisante pour estimer un nombre important de gaussiennes par densité. Après 4 divisions, le nombre moyen de gaussiennes par densité est d'environ 8 au lieu de 16 théoriquement. Pour les trois autres corpus (chiffres, Trégor et nombres) le nombre moyen de gaussiennes par densité est de l'ordre de 15 après 4 divisions.

COMPARAISON DES APPROCHES

La figure ci-dessous illustre la différence conceptuelle entre un monomodèle à densités multigaussiennes et un multimodèle.

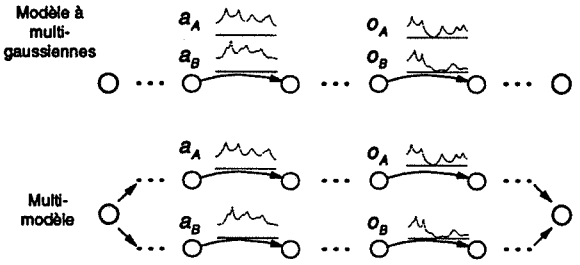


Figure 6 - Répartition des paramètres entre un modèle à multigaussiennes et un multimodèle.

Dans le cas d'un modèle à multigaussiennes, on peut avoir indépendamment un spectre de type a_A ou a_B à un instant donné, et plus loin un spectre de type o_A ou o_B. Dans le cas d'un multimodèle, on aura soit un spectre de type a_A suivi d'un spectre de type o_A, ou un spectre de type a_B suivi d'un spectre de type o_B. L'approche par multimodèles prend donc en compte une certaine corrélation entre les modélisations des différents sons. La prise en compte de cette corrélation devrait normalement favoriser la discrimination entre phrases, et ce d'autant plus que les phrases sont longues : plus la phrase est longue, plus il y a de sons, et donc plus la prise en compte d'une corrélation entre les sons devrait améliorer les performances de reconnaissance.

La figure suivante indique les taux d'erreurs obtenus sur les différents corpus (partie test) en fonction du nombre total de paramètres des modèles. Le but est d'évaluer l'apport des multimodèles et des multigaussiennes. Des échelles logarithmiques sont employées sur chaque axe. Les bimodèles utilisent une classification a priori (hommes/femmes) des locuteurs

lors de la phase d'apprentissage. Aucune information concernant le locuteur n'est employée en phase de reconnaissance, les deux modèles sont disponibles, celui conduisant à la plus forte probabilité d'émission impose sa réponse. Pour le monomodèle et pour le bimodèle on a augmenté progressivement le nombre de gaussiennes de chacun des modèles par divisions successives, puis apprentissage, comme décrit précédemment. Cette figure permet donc de comparer aisément les performances d'un bimodèle à N gaussiennes par densité, avec celles d'un monomodèle ayant 2N gaussiennes par densité.

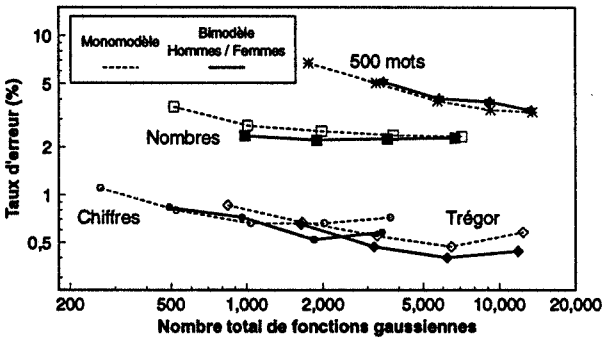


Figure 7 - Comparaison des performances sur les données brutes (i.e. sans adaptation aux lignes).

Cette figure, qui présente les résultats obtenus sur les données brutes, fait ressortir un avantage de l'approche par multimodélisation pour la reconnaissance indépendante du locuteur sur le corpus des chiffres et le corpus Trégor. Pour le corpus des nombres, l'avantage est moins net, et il s'atténue avec l'augmentation du nombre de gaussiennes par densité. Quant au corpus des 500 mots, les résultats sont sensiblement équivalents entre les deux approches (le bimodèle étant légèrement moins bon).

CONCLUSION

Dans ce papier nous avons comparé diverses modélisations markoviennes de la parole en nous intéressant principalement aux structures des modèles et à la classification des locuteurs. De manière globale, toute augmentation progressive du nombre de paramètres conduit à une amélioration des performances de reconnaissance.

L'augmentation progressive du nombre de modèles et l'application d'une procédure de classification automatique des locuteurs couplée à l'apprentissage des paramètres semble prometteuse. Elle permet de s'affranchir de la classique partition hommes/femmes et d'aller au delà de deux classes.

L'augmentation progressive du nombre de gaussiennes des densités continues améliore également de manière significative les performances de reconnaissance. L'amélioration beaucoup plus importante sur les données d'apprentissage que de test est révélatrice de la taille limitée des bases de données employées dans cette étude.

La stabilité des procédures reposant sur une augmentation progressive et contrôlée des paramètres des modèles est à noter. Toutes les procédures travaillant par division des classes (locuteurs associés à une branche ou trames associés à une densité) n'ont jamais conduit à une dégradation des performances.

REMERCIEMENTS

Les auteurs tiennent à remercier J. MONNE et C. MOKBEL pour les nombreuses discussions relatives à la modélisation markovienne et à l'adaptation aux lignes téléphoniques, ainsi que pour la relecture de ce papier.

BIBLIOGRAPHIE

[1] L. R. Rabiner, J. G. Wilpon, B. H. Juang : "A performance evaluation of a connected digit recognizer" ; *Proc. IEEE Int. Conf. ASSP*, Dallas, Texas, USA, avril 1987, pp. 101-104.

[2] L. R. Rabiner, C. H. Lee, B. H. Juang & J. G. Wilpon : "HMM clustering for connected word recognition" ; *Proc. IEEE Int. Conf. ASSP*, Galsgow, Ecosse, mai 1989, pp. 405-408.

[3] D. Jouvet : *"Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques"* ; Thèse de doctorat de l'ENST, Paris, juin 1988.

[4] D. Van Compernolle, J. Smolders, P. Jaspers & H. Hellemans : "Speaker clustering for dialectic robustness in speaker independent recognition" ; *Proc. EUROSPEECH*, Gènes, Italie, septembre 1991, pp. 723-726

[5] L. Mathan & L. Miclet : "Speaker hierarchical clustering for improved speaker-independent HMM word recognition" ; *Proc. Int. Conf. ASSP*, Albuquerque, Nouveau-Mexique, USA, avril 1990, pp. 149-152.

[6] B. H. Juang & L. R. Rabiner : "Mixture autoregressive hidden Markov models for speaker independent isolated word recognition" ; *Proc. IEEE Int. Conf. ASSP*, Tokyo, Japon, avril 1986, pp. 41-44.

[7] H. Ney : "Acoustic modelling of phonemes units for continuous speech recognition" ; *Signal Processing V : Theories and Applications*, L. Torres, E. Masgrau & M. A. Lagunas (eds.), Elsevier Science Publishers, B.V., 1990, pp. 65-72.

[8] D. Jouvet, L. Mauuary & J. Monné : "Automatic adjustments of the Markov models topology for speech recognition applications over the telephone" ; séminaire NATO ASI de juillet 1990, Cetraro, Italie, publié en 1992 dans *Speech Recognition and Understanding, Recent Advances*, NATO ASI Series, Vol. F 75, pp. 65-70.

[9] C. Gagnoulet : "Speech recognition over the telephone : experiments in France" ; *Proc. Voice Systems Worldwide 1990 conference*, Londres, mai 1990, pp. 173-177.

[10] D. Jouvet, K. Bartkova & J. Monné : "On the modelization of allophones in an HMM based speech recognition system" ; *Proc. EUROSPEECH*, Gènes, Italie, septembre 1991, pp. 923-926.

[11] K. Bartkova & D. Jouvet : "Modelization of allophones in a speech recognition system" ; *Proc. Int. Congress on Phonetic Sciences*, Aix en Provence, France, août 1991, Vol. 4, pp. 474-477.

[12] C. Mokbel, J. Monné & D. Jouvet : "On-line adaptation of a speech recognizer to variations in telephone line conditions"; *Proc. EUROSPEECH*, Berlin, Allemagne, septembre 1993, pp. 1247-1250.

[13] D. Jouvet, L. Mauuary & J. Monné : "Automatic adjustments of the structure of Markov models for speech recognition applications" ; *Proc. EUROSPEECH*, Gènes, Italie, septembre 1991, pp. 927-930.

UNE METHODE TEMPORELLE DE DISCRIMINATION BRUIT / PAROLE EN MILIEU BRUITE

Jean-Baptiste PUEL, Régine ANDRE-OBRECHT.

IRIT - URA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex

Résumé

La détection des début et fin de parole est un traitement stratégique dans un système de reconnaissance automatique de la parole en milieu bruité. Nous proposons une méthode de détection qui conjugue la segmentation statistique du signal et l'analyse du signal temporel. Des expériences sont faites sur la base de données enregistrée au cours du projet ESPRIT ARS : des mots isolés ont été prononcés dans une voiture aux vitesses de 130 km/h et 90 km/h.

Les résultats montrent qu'une parfaite discrimination entre segment "bruit" et segment "parole" est obtenue; de plus, une meilleure précision est atteinte quant à la localisation de ces début et fin de parole.

Mots clés : *segmentation, reconnaissance automatique , détection bruit/parole*

1. INTRODUCTION

A l'heure actuelle, de nombreux systèmes de reconnaissance automatique de parole (RAP) obtiennent dans le cadre d'applications simple (mots isolés-monolocuteur) des résultats excellents en ambiance dite propre (type laboratoire). Malheureusement une dégradation drastique des performances se produit en conditions réelles, dès que le milieu est dit "bruité" [Juang 91]. Le bruit ambiant provoque non seulement une distorsion additive du signal, mais il est à l'origine de modifications au niveau de la production même des sons (effet Lombard [Lombard 11]).

Une des premières manières de rendre plus robustes les systèmes de RAP, est de détecter l'apparition et la disparition de la parole dans le signal : la localisation du bruit seul permettra une meilleure adaptation des modèles de bruit en cours de reconnaissance ; la localisation correcte de la parole rendra, en phase d'apprentissage, les lois d'observation plus discriminantes et, en phase de reconnaissance, elle conduira à un calcul des vraisemblances plus optimal. Sujet apparemment simple, il reste délicat à traiter selon les connaissance que l'on a du bruit a priori, il a fait l'objet de nombreux travaux [Lamel 81], [Junqua 92].

Les méthodes sont en général basées sur des calculs de paramètres classiques tels que l'énergie du signal, son énergie résiduelle, en se limitant à des bandes de fréquence précises. Des automates à seuils régissent les décisions et un relâchement aveugle des frontières est systématiquement appliqué pour éviter de tronquer les mots. La méthode que nous proposons est basée sur une analyse temporelle très simple, sa force repose sur le couplage de l'analyse à un algorithme de segmentation automatique performant. La méthode de "Divergence forward-backward" permet en effet de localiser les zones quasi-stationnaires du signal sans toutefois les identifier; en particulier les frontières bruit/parole et parole/bruit sont correctement détectées mais ne sont pas identifiées.

Cette méthode a été expérimentée sur un corpus formé de mots isolés prononcés dans une voiture roulant à 90km/h et 130 km/h.

2. ANALYSE TEMPORELLE

Même si l'énergie du signal de parole n'est pas un paramètre robuste en milieu bruité, et même si le rapport signal/bruit descend à 0dB, il n'en demeure pas moins que les noyaux vocaliques correspondent toujours à des maxima d'amplitude du signal.

L'idée de la détection temporelle est d'analyser le signal en cumulant les différences des valeurs d'amplitude successives. Ce cumul est une approximation de l'abcisse curviligne, puisque la discrétisation selon la coordonnée temporelle est constante et correspond au pas d'échantillonage. Nous ferons par la suite cet abus de langage.

Nous faisons l'hypothèse que l'abcisse curviligne doit accuser une croissance plus rapide dans les zones de parole de type voisé que dans les zones de bruit.

Pour quantifier ce phénomène, nous introduisons les paramètres suivants :

si
- s(t) représente l'abcisse curviligne du signal de parole y(t).
- t désigne l'indice temporel et L un nombre d'échantillons fixé a priori,

nous posons :
- $S(n) = s(nL) - s((n-1)L)$
- $DS(n) = S(n) - S(n-1)$

S(n) représente une valeur moyenne de la "longueur de la courbe" par unité de temps (ou trame) et DS(n) sa dérivée.

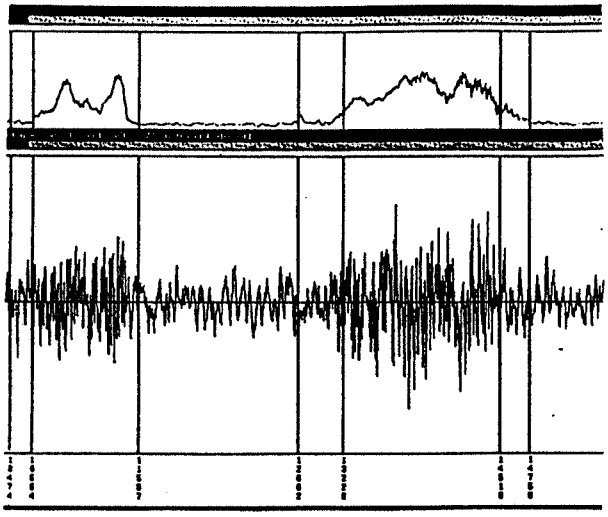


Figure I : Signal de parole correspondant à la pronciation du nom "Lipka" et fonction S correspondante.

Sous l'hypothèse que le bruit ambiant est stationnaire par morceaux, la fonction S doit peu varier dans les zones de bruit, croître rapidement en début de parole et décroître en fin de parole ; en d'autres termes, la fonction DS présentera une valeur très faible dans les zones bruitées et verra sa valeur augmenter brutalement en début de parole, puis décroitre pour retrouver son niveau habituel dans le bruit. Ainsi, DS, par ses brusques variations nous indiquera l'apparition et la disparition de la parole dans le signal.

Afin de détecter ces variations, nous avons dans un premier temps introduit un seuil λ :

si $DS(n) < \lambda$, la trame analysée correspond à du bruit.

si $DS(n) \geq \lambda$, la trame correspond à de la parole bruitée.

Cependant, l'expérience montre que cette façon de discriminer le bruit de la parole est trop "brutale", et peut fournir des résultats erronés lorsque le niveau du bruit est élevé.

Pour affiner les résultats, nous avons utilisé deux seuils λ_1 et λ_2.

Si $DS(n) < \lambda_1$, la trame n correspond à du bruit.

Si $DS(n) > \lambda_2$, la trame correspond à de la parole bruitée.

Enfin, pour $\lambda_1 < DS(n) < \lambda_2$, la trame n correspond soit à une zone de transition bruit-parole ou parole-bruit, soit à une trame de bruit non stationnaire, soit à une trame de silence lors de la prononciation d'un mot.

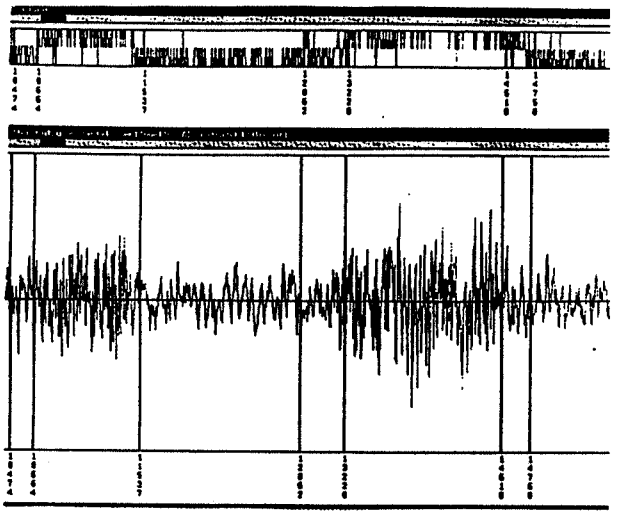


Figure II : Signal de parole correspondant à la pronciation du nom "Lipka" et fonction DS seuillée correspondante.

3. COUPLAGE DE L'ANALYSE TEMPORELLE ET DE LA SEGMENTATION

L'algorithme de segmentation automatique est la méthode de "Divergence forward-backward" [André-Obrecht 88]. Il permet de localiser les zones de signal quasi-stationnaires, ses performances sont indépendantes du corpus, du locuteur et surtout des conditions d'enregistrement. Des expériences ont été réalisées sur des signaux enregistrés en voiture avec un micro omnidirectionnel, les résultats sont très satisfaisants [André-Obrecht 93]. Nous avons en particulier voulu utiliser le fait que les débuts et fins de parole sont correctement détectés, mais les frontières ne sont pas identifiées comme telles.

3.1. DISCRIMINATION PAROLE - BRUIT

Chacun des segments est parcouru de façon à calculer les valeurs de l'abcisse curviligne moyenne par trame ainsi que sa dérivée (S(n) et DS(n)).

Les valeurs prises par DS par rapport aux seuils λ_1 et λ_2 calculés automatiquement vont permettre de discriminer les segments de la façon suivante :

pour tout segment $[t_o, t_f]$:

• si la moyenne des valeurs prises par DS(n) pour n vérifiant $t_o < nL < t_f$, est inférieure à λ_1, le segment est classé "bruit".

• si cette moyenne est supérieure à λ_2, le segment est classé "parole".

• dans tous les autres cas, le segment est classé en fonction des segments adjacents et de leur durée.

Si le segment d'indice i présente une valeur moyenne de DS comprise entre λ_1 et λ_2, ce segment sera classé "parole" si le segment d'indice i-1 ou le segment d'indice i+1 est un segment de parole et si sa durée est inférieure à d_{max}. Pour les segments en début et fin de parole, cette opération est synonyme d'un relâchement de frontières.

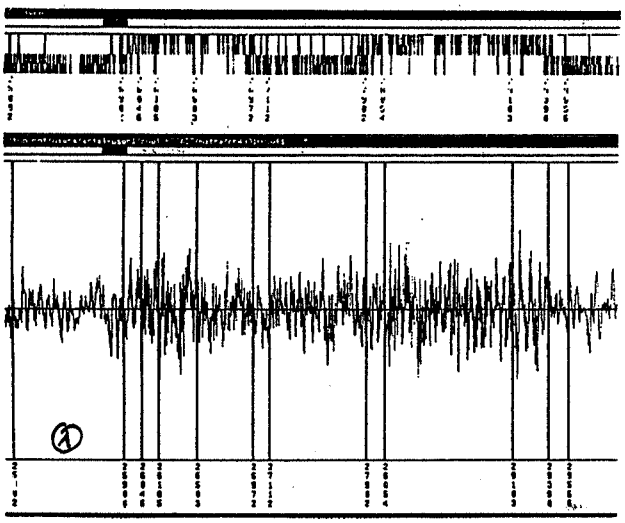


Figure III : exemple de relachement provoqué par la segmentation.

Le segment n°1, pour lequel la valeur de DS est comprise entre λ_1 et λ_2, est associé au mot adjacent.

3.2. REGLAGE DES PARAMETRES

Les seuils λ_1 et λ_2 sont déterminés automatiquement à partir de la valeur moyenne de DS calculée sur les premières trames de signal considérées comme du bruit.

Ces seuils sont donc indépendants du bruit, mais nécessitent une réelle ambiance bruitée.

Le seuil, d_{max}, durée maximale pour laquelle un segment douteux est accolé au mot voisin est une contrainte imposée par le système. Une valeur trop élevée permettra de conserver un maximum d'information, mais risque de provoquer la reconnaissance d'un seul mot alors que deux mots avaient été prononcés. Une valeur faible de d_{max} présentera les avantages et inconvénients inverses : peu de risque de concaténer deux mots, mais risque de perdre le premier ou le dernier segment du mot reconnu. Cette solution présente de plus l'inconvénient de scinder éventuellement en deux un mot prononcé lentement.

La solution retenue : donner à d_{max} une valeur assez élevée (de l'ordre de 250 à 400 ms) s'avère satisfaisante à l'expérimentation. Si l'on veut coupler ce traitement à un système de reconnaissance de mots isolés, cette valeur contraint le locuteur à respecter une coupure franche entre deux mots consécutifs (ce qui interdit la parole continue), à moins d'autoriser la reconnaissance de mots connectés.

La valeur de L a été fixée à 4ms, soit 32 échantillons pour une fréquence d'échantillonnage de 8 kHz.

4. EXPERIMENTATION - RESULTATS

Comme nous l'avons dit précédemment, la méthode a été testée sur une base de données enregistrées au cours du projet Esprit ARS : ce corpus est formé de 43 mots isolés (noms propres), prononcés 8 fois par chacun des quatre locuteurs dans une voiture. Quatre répétitions sont prises à 130 km/h et quatre à 90km/h. Le rapport signal/bruit est compris entre 0 et 30 db, la valeur moyenne est de 15 db. Une segmentation manuelle est fournie avec les données afin de réaliser l'apprentissage de modèles de références pour un système de reconnaissance ; dans la mesure où les performances du système seraient fortement dégradées si des mots étaient tronqués, ces frontières sont fortement relâchées.

Les résultats sont comparés à cette segmentation manuelle du corpus. Les différences observées ne sont pas à proprement parler des erreurs, mais résultent de la décision prise lors de la segmentation manuelle de relâcher fortement les frontières de façon à ne pas tronquer les mots.

	130 km/h	90 km/h	Total
Locuteur 1			
Début	510	479	498
Fin	876	1734	1219
Locuteur 2			
Début	391	419	402
Fin	777	1650	1246
Locuteur 3			
Début	427	447	435
Fin	885	1997	1330
Locuteur 4			
Début	628	725	652
Fin	728	1023	802

Tableau I : erreur moyenne par locuteur en début et fin de mot (en nombre d'échantillons).

Ainsi les frontières de début de mot déterminées automatiquement sont-elles toujours à l'intérieur des frontières manuelles, avec une différence moyenne de 496 échantillons, soit environ 62 ms.

Quant aux frontières automatiques de fin de mot, elles sont éloignées des frontières manuelles de 1149 échantillons en moyenne, soit environ 142 ms.

Sur ce tableau, on constate que l'erreur moyenne se situe entre 400 et 2000 échantillons, soit 50 ms à 250 ms. On remarque que les résultats sont meilleurs à 130 km/h, où le niveau de bruit est plus élevé : cette méthode présente ses meilleurs résultats dans une ambiance réellement bruitée.

De plus, les résultats sont plus proches de la segmentation manuelle en début de mot qu'en fin de mot.

On remarque d'ailleurs qu'en fin de mots, la frontière automatique n'est pas systématiquement située avant la frontière manuelle : la façon de relacher les bornes est différente.

	AVANT		APRES	
	Err. moy.	Freq. (%)	Err. moy.	Freq. (%)
Locuteur 1	341	24	1352	76
Locuteur 2	618	20	1318	80
Locuteur 3	1141	17	1365	83
Locuteur 4	407	52	1276	48

Tableau II : position des frontières de fin de mot.

5. CONCLUSION

Ces résultats se sont avérés encourageants, aussi, nous poursuivons cette étude de la manière suivante :

• grâce à une segmentation manuelle encore plus précise et détaillée, nous étudions quels sons sont acoustiquement présents et effectivement tronqués.

• nous comparons ce detecteur avec un détecteur basé sur la dérivée spectrale, nous l'intégrerons éventuellement à notre système.

• nous réalisons un système de reconnaissance automatique de la parole utilisant comme pré-traitement
- la segmentation automatique du signal.
- la détection bruit/parole
- un débruitage de type NSS [Lockwood 92].

Ce système sera comparé aux approches classiques qui n'utilisent qu'une détection bruit/parole grossière : cette comparaison permettra de mesurer les performances d'une système évolué de détection en RAP.

REFERENCES BIBLIOGRAPHIQUES

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, vol. 36 pp 26-40, January 1988.

[André-Obrecht 93] R. André-Obrecht, "Segmentation et parole ?", Habilitation à diriger des recherches, Université de Rennes I, Juin 1993.

[Juang 91] B.H. Juang, "Speech recognition in adverse environment", Computer Speech and Language, vol. 5, pp 275-294, 1991.

[Junqua 92] B.Mak, J.C. Junqua, B. Reaves, "A robust speech/non-speech detection algorithm using time and frequency-based features", ICASSP 1992.

[Lamel 81] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. on ASSP, vol. 29, pp 777-785, August 1981.

[Lockwood 92] P. Lockwood, J. Boudy, M. Blanchet, "Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environment", ICASSP 1992.

[Lombard 11] Lombard, "Le signe de l'élévation de la voix", ANN. Maladies oreille, larynx, nez, pharynx, 37, pp 101-119, 1911.

DEUX SYSTEMES D'ETIQUETAGE PHONETIQUE BASES SUR UNE PRE-SEGMENTATION DU SIGNAL DE PAROLE

A. FARHAT, N. PARLANGEAU, R. ANDRE-OBRECHT

IRIT - URA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex

Résumé

Nous présentons dans cette communication deux systèmes d'étiquetage phonétique basés sur la pré-segmentation du signal et la modélisation Markovienne des unités élémentaires. Leur segmentation diffèrent sur le type de paramètres utilisé (fréquentiel, temporel) et le type d'approche (statistique, à base de connaissance). Les résultats obtenus par chaque système sur le corpus français d'EUROM0 sont présentés.

Mots clés: *étiquetage phonétique, segmentation, événement phonétique, modèles de Markov cachés*

1. INTRODUCTION

Quel que soit le domaine de recherche en traitement automatique de la parole, il faut disposer de corpus de parole étiquetés et annotés de manière à pouvoir consulter les différents constituants de ces bases. Pour rendre plus efficace la phase d'apprentissage d'un système de reconnaissance de la parole, il est nécessaire d'initialiser les paramètres à l'aide d'un étiquetage fin des données.

L'étiquetage d'une partie de la base de données de parole est généralement considérée comme utile pour quantifier ou apprendre les procédés automatiques. Compte tenu de la dimension des corpus utilisés, actuellement on fait de plus en plus appel à l'étiquetage automatique. Deux critères majeurs ont induit l'automatisation des systèmes d'alignement : le temps et la robustesse (ou cohérence) dans le procédé. Un étiqueteur manuel entraîné met environ une minute pour étiqueter un phonème [Boë-88]. Par ailleurs les évaluations de l'étiquetage manuel ont montré que l'étiqueteur manuel n'est pas toujours constant dans l'utilisation —choix et hiérarchie— des critères de segmentation [Leung 84] et qu'il est de plus difficile d'élaborer des critères robustes utilisables par un groupe d'étiqueteurs humains.

Depuis de nombreuses années, nous nous intéressons au problème de l'alignement automatique du signal de parole sur sa transcription phonétique. Nos travaux nous ont conduits à proposer deux systèmes d'étiquetage phonétique, le système DIVERGENCE et le système SALAM. Leur originalité principale provient de l'utilisation au niveau du pré-traitement acoustique, d'une procédure de segmentation. Par contre, les deux algorithmes de segmentation diffèrent de par le type de paramètres utilisés (fréquentiels et temporels) et de par la procédure elle même : le système DIVERGENCE relève d'une étude statistique du signal (détection séquentielle de rupture) ; le système SALAM est un système basé sur des connaissances et la segmentation résulte d'un étiquetage centiseconde. Des comparaisons entre les deux procédures de segmentation ont été effectuées [André-Obrecht-92] ; les segments obtenus sont infra-phonétiques et le taux de sur-segmentation qui correspond au nombre de segments par son, est de l'ordre de 2. Le système basé sur des connaissances a donné naissance à de nombreux systèmes d'étiquetage [Farhat-92], mais un seul est présenté dans cette communication.

A la suite du pré-traitement segmental, l'alignement du signal sur la transcription phonétique auditive est réalisé à l'aide de modèles de Markov cachés. Le système DIVERGENCE utilise des modèles de type discret et une phase de quantification des paramètres s'avère donc nécessaire, tandis que le système SALAM utilise des modèles continus.

Les expériences que nous décrivons ont été réalisées sur le corpus français d'EUROM0. Le système DIVERGENCE n'en étant qu'à ses premiers développements, est dépendant du locuteur à la différence du système SALAM qui est une version améliorée de précédents systèmes et qui est indépendant du locuteur.[Farhat 93].

2. SYSTEME DIVERGENCE

Dans bon nombre de systèmes de reconnaissance automatique de parole dits analytiques comme dans les systèmes d'étiquetage automatique, une première étape consiste généralement à transcrire le signal acoustique en une suite d'unités élémentaires. Cette étape appelée décodage acoustico-phonétique se décompose en trois phases, la paramétrisation du signal, la segmentation et l'idenfication ; très souvent la phase de segmentation est inexistante ou implicite. Le système DIVERGENCE s'appuie sur une segmentation statistique *a priori*. Chaque segment est ensuite étiqueté à l'aide d'un dictionnaire de références obtenu par quantification vectorielle ; l'alignement final est obtenu par modélisation Markovienne.

2.1 LE PRETRAITEMENT ACOUSTIQUE

Comme nous l'avons déjà dit, la spécificité de ces approches réside dans l'utilisation de la segmentation comme pré-traitement. La méthode utilisée ici, à savoir la méthode de divergence "forward-backward", est basée

sur une modélisation statistique du signal et un test statistique [André-Obrecht 88]. Les segments obtenus sont infra-phonétiques et correspondent aux parties stables et transitoires des sons, mais aucune connaissance sur leur nature n'est utilisée pour guider éventuellement une paramétrisation spécifique.

Chaque segment est paramétrisé par huit coefficients cepstraux (MFCC) et l'énergie. Ces paramètres sont calculés sur une fenêtre de 256 échantillons centrée sur le milieu du segment après une transformée de Fourier.

2.2 L'IDENTIFICATION

Chaque segment est étiqueté phonétiquement. Dans ce but, nous avons regroupé les sons du français selon leurs principales caractéristiques en dix classes (tableau 1). Les étiquettes correspondent à ces dix classes et pour chacune d'elles, nous avons construit un sous dictionnaire de références.

Etiquettes	Signification
K	Noyau vocalique central
M	Noyau vocalique grave
N	Noyau vocalique aigu
L	Occlusive/Liquide/Nasale htes freq.
U	Occlusive/Liquide/Nasale basses freq.
O	Barre de voisement
Z	Fricative voisée
S	Fricatve non voisée
X	Silence bruité
Q	Silence

Tableau 1: Les classes phonétiques

La construction d'un dictionnaire de références est en général une étape critique ; la qualité d'une telle méthode repose d'une part sur le choix des paramètres et d'autre part sur le choix de l'algorithme de construction lui-même. Nous avons choisi d'utiliser une méthode de type Quantification Vectorielle (VQ) ; de nombreuses études ont été faites sur ce sujet, nous avons opté pour une version améliorée de l'algorithme L.B.G. [Su 90].
L'ensemble d'apprentissage est décomposé en dix sous ensembles d'apprentissage correspondant aux dix classes; l'algorithme L.B.G. modifié est appliqué à chaque sous ensemble et permet d'obtenir un sous-dictionnaire par classe. Le dictionnaire de références est obtenu par réunion de ces sous-dictionnaires et il s'en suit que chaque référence appartenant à un unique sous-dictionnaire est étiquetée de la classe phonétique correspondante.
Lors de l'identification des segments, chaque vecteur de paramètres quéfrentiels est comparé à l'ensemble des références du dictionnaire global à l'aide de la règle du plus proche voisin (distance euclidienne) ; est associée à ce segment l'étiquette de la référence la plus proche.
Il est évident que la connaissance des débuts et fins de parole sont des points d'ancrage importants dans un système d'alignement automatique. Or nous traitons des phrases continues sans projection des pauses ; il nous faut donc rendre plus robuste cette détection. Nous identifions chaque segment comme du silence en fonction de son taux de passages par zéro et de son énergie et nous corrigeons en conséquence l'étiquette phonétique obtenue par QV.

2.3 L'ALIGNEMENT PHONETIQUE

L'alignement phonétique du signal sur la transcription auditive de l'expert est basé sur une modélisation Markovienne discrète. Nous avons défini quatorze unités de base qui correspondent à quatorze modèles de Markov cachés élémentaires (Tableau 2). L'unité de base doit être représentative à la fois d'un type de son et de sa représentation segmentale. Ainsi, nous avons été amenées à proposer une unité de type explosion d'une plosive, une unité de type silence de plosive. Certains sons ont pu être regroupés : segmentalement ils se décomposent de manière semblable, l'identification des segments correspondants donnent les mêmes étiquettes phonétiques conduisant à des lois de probabilité d'observation identiques.

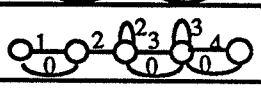
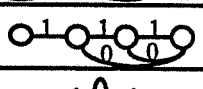
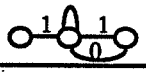
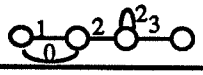
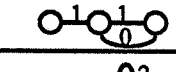
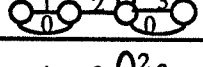
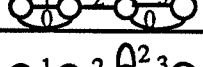
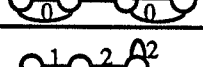
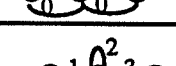
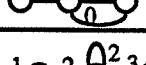
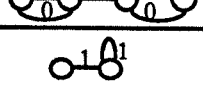
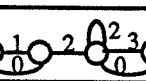
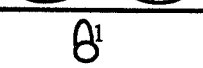
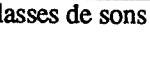
Unités de base	Groupes de sons	Topologie des modèles
A	A,E,I,O,U,Y,e *voyelles*	
AN	AN,IN,ON *voyelles nasales*	
B	B,D,G *plosives voisées*	
Bb	Bb,Db,Gb *explosion*	
P	P,T,K *plosives sourdes*	
Pb	Pb,Tb,Kb *explosion*	
F	F,S,CH *fricatives sourdes*	
Z	Z,z *fricatives voisées*	
L	M,N,L *liquides*	
w	w,y *semi-voyelles*	
j	j *semi-voyelle*	
V	V *liquide voisée*	
R	R *non voisé*	
r	R *voisé*	
Q	*Silence*	

Tableau 2 : M.M.C des classes de sons

La procédure d'alignement est la suivante : pour chaque phrase, la description auditive en phonèmes fournie par l'expert phonéticien est transcrite à l'aide de ces unités de base. A partir des modèles de Markov élémentaires correspondant, un réseau global est obtenu et est modifié à l'aide de règles de coarticulation. Ces règles sont liées à

- des phénomèmes d'assimilation (rl -> l),
- des phénomènes de coarticulation forte (f s -> f ou s)
- des problèmes d'omissions dûes à la segmentation comme l'omission de l'explosion d'une plosive sonore.

La suite d'étiquettes phonétiques est présentée à ce réseau modifié, l'alignement s'effectue à l'aide de l'algorithme de Viterbi.

3. SYSTEME SALAM

Le système SALAM (Système d'ALignement Automatique basé sur les modèles de Markov cachés) décrit dans cette communication est une extension des travaux que nous menons depuis plusieurs années sur l'étiquetage phonétique automatique. Ce système comporte deux composantes principales :

- le module de segmentation et d'étiquetage en événements phonétiques appelé SED (Segmentation, Etiquetage et Discontinuité) ;
- le module d'alignement des événements phonétiques sur la transcription phonétique. Pour les expériences décrites dans cette communication, la transcription phonétique alignée est fournie par l'expert phonéticien.

3.1. PRE-TRAITEMENT ACOUSTIQUE : LE SYSTEME SED

Le système SED a été développé dans le cadre du projet ESPRIT SAM (Speech input/output Assessment and Methodology) n° 2589 [Pérennou-91]. Il est basé sur une analyse temporelle du signal de parole. Deux paramètres, l'amplitude et le taux de passage à zéro, leurs dérivées premières et secondes, et un ensemble de règles acoustico-phonétiques vont activer l'étiquetage phonétique des trames de 4 ms. Par la suite, les trames successives ayant la même étiquette sont reproupées et les événements isolés ou transitoires sont lissés. Le résultat du module SED est donc une segmentation et un étiquetage en événements phonétiques.

3.2. L'ALIGNEMENT PHONETIQUE

Pour produire un étiquetage "phonétique", il faut insérer après la procédure de segmentation et de paramétrisation, une étape de modélisation des unités phonétique en terme d'événements. Comme dans la précédente méthode, nous avons choisi la modélisation Markovienne. L'alignement est effectué moyennant la construction du réseau de Markov associé à la transcription phonétique auditive et l'application de l'algorithme de Viterbi.

Dans la première série d'expériences que nous avons effectuées, nous avons utilisé comme observations à aligner, les étiquettes phonétiques associées aux segments (observations de type discrets). Pour ces expériences, nous avons regroupé les différentes unités phonétiques en dix classes majeures, à chaque classe était associé un modèle de Markov cachés discrets [Farhat-92]. Nous avons amélioré sensiblement les résultats avec l'introduction d'une procédure de lissage des probabilités d'observation.

Par la suite une question s'est posée naturellement, à savoir si l'on extrait une autre information des segments apporte-t-on des améliorations à la procédure d'étiquetage dans sa globalité ? Pour répondre à cette question nous avons procédé à une deuxième série d'expérimentations où les observations à aligner sont des vecteurs de coefficients cepstraux associés aux segments. Pour ces expériences, il a été nécessaire de définir un plus grand nombre de classes phonétiques (dix-neuf) [Farhat-93]. Au cours de ces expérimentations le taux d'erreurs graves du système (à ±50 ms de l'étiquetage manuel) était de l'ordre de 10%. Ces erreurs graves en alignement sont dûes à la sous-segmentation du pré-traitement acoustique.

Les travaux que nous présentons concernent les améliorations que nous avons apportées à cette deuxième version du système SALAM utilisant des modèles de Markov cachés continus. Ces améliorations portent sur trois points essentiels :

- le nombre de classes d'unités phonétiques ;
- une solution au problème de frontières manquantes pendant la phase de segmentation ;
- l'adaptation non supervisée à un nouveau locuteur.

Nous avons effectué de nouvelles expériences avec 23 classes phonétiques. Cette nouvelle classification est donnée par le tableau 3. Les modèles de Markov associés aux classes phonétiques sont identiques à ceux décrits dans [Farhat-93].

classes phonétiques	sons	classes phonétiques	sons
A	a ɑ	O	o ɔ
E	e ɛ	H	ø œ
I	i y	U	u
Y	ə	-	-
C	ɔ̃ ɑ̃	M	œ̃ ɛ̃
J	j ɥ	W	w
N	m n ɲ	L	l
R	r voisé	X	r non voisé
P	p	T	t
K	k	B	b d g
S	s ʃ	F	f
Z	z ʒ	V	v

Tableau 3 : Classification des différentes sons du français pour le système SALAM2-23.

Cette nouvelle classification a permis d'améliorer légèrement les résultats de notre système.

3.3. TRAITEMENT DE LA SOUS-SEGMENTATION

Un problème majeur pour tout système d'étiquetage ayant comme pré-traitement acoustique la segmentation du signal, réside dans le fait que les erreurs de segmentation sont répercutées pendant la phase d'alignement. Chacune de ces erreurs engendre souvent plusieurs erreurs d'alignement. Dans nos expériences passées nous avions tenté de remédier à ce problème en autorisant le saut de certaines unités phonétiques. Dans ce but, des statistiques étaient effectuées sur le corpus d'apprentissage pour identifier les unités phonétiques dont la frontière n'était pas détectée pendant la segmentation. Cette démarche n'ayant pas été concluante au niveau des résultats "quantitatifs", nous l'avions abandonnée par la suite.

Au cours de la démarche précédente, nous avons tenté de résoudre le problème de sous-segmentation pendant la phase d'alignement sans remettre en cause la segmentation même. Une autre approche consiste à résoudre le problème à sa source, au niveau des segments acoustiques. Pour cela il faut pouvoir identifier et localiser les segments regroupant deux ou plusieurs unités phonétiques.

Notre première démarche a consisté en une division simple des segments "trop longs". Dans la mesure où les erreurs de sous-segmentation correspondent à des segments de durée au moins égale à 80 ms, nous avons donc décomposé ces segments en plusieurs sous-segments (de durée identique). Une série d'expériences a été menée pour déterminer la durée minimum des segments décomposés et définir le nombre de sous-segments. Les meilleurs résultats ont été obtenus avec une durée minimum de segment de 100 ms et une décomposition en 3 sous-segments.

Il est clair que cette méthode est une méthode aveugle, aucun traitement sur ces segments n'est effectué. Au regard des performances ainsi obtenues, l'insertion d'un post-traitement "plus intelligent" pour détecter les frontières manquantes au sein des segments "trop longs" ne peut qu'améliorer ces résultats.

3.4. ADAPTATION NON SUPERVISEE A UN NOUVEAU LOCUTEUR

Pour faire face à la taille réduite du corpus utilisé en phase d'apprentissage et au nombre élevé de modèles à apprendre, nous proposons une adaptation automatique de notre système à un nouveau locuteur, voire un nouveau corpus.

Nous avons inséré, avant la phase d'alignement, un processus d'adaptation automatique aux données de test. Cette opération permet d'ajuster les modèles suivant les données de test et d'assurer une meilleure adaptation du système de la manière suivante :

Après un premier alignement du corpus de test, les probabilités d'observation et de transition des modèles sont réajustées. Nous ne supervisons pas cette étape de sorte que les erreurs d'alignement ne sont pas corrigées. Enfin, l'alignement proprement dit est effectué en utilisant ces modèles adaptés.

Avec ce procédé, nous avons amélioré les performances du système d'étiquetage tout en préservant sa souplesse d'exécution. Cette adaptation non supervisée facilite l'utilisation de notre système sur de nouveaux corpus tels que BREF. Les modèles de base sont enrichis avec les nouvelles données étiquetées automatiquement. Bien qu'un certain nombre d'erreurs fassent partie de ces données, les performances du système sont augmentées par cette méthode. Le processus d'adaptation non supervisée semble être une étape importante dans les applications futures de notre système.

4. RESULTATS RESPECTIFS

4.1 SYSTEME DIVERGENCE

Les expériences effectuées sont monolocuteur, mais nous rappelons que le pré-traitement acoustique n'a pas nécessité d'apprentissage.

Le corpus d'apprentissage est constitué des dix premières phrases d'un locuteur de la base EUROM0.
Dans un premier temps, nous avons manuellemnt étiqueté les segments en étiquettes phonétiques, afin de construire le dictionnaire de références.
Dans un second temps, nous avons projeté l'étiquetage manuel fourni par EUROM0 sur les segments ; les observations nécessaires à l'apprentissage des M.M.C. élémentaires sont les étiquettes phonétiques obtenues après application du dictionnaire sur l'ensemble d'apprentissage segmenté.

Les expérimentations effectuées portent d'une part sur le système Divergence sans post-traitement silence/parole des segments (Divergence1) ; nous distinguons les résultats selon l'utilisation d'une projection des pauses ou non. Une autre série d'expérimentations a été effectuée avec le système décrit plus avant c'est-à-dire avec post-traitement et sans projection des pauses.
Dans les deux cas, l'ensemble test est constitué des treizes dernières phrases du même locuteur qu'en apprentissage.

	Corrects ±20 ms (%)	Erreurs ±20 ms (%)	Omissions
Divergence1 ss post-trait ss pauses	71	27	2
Divergence1 avec pauses	73	24	2
Divergence2 post-trait ss pauses	74	23	2

Tableau 4 : Résultats obtenus avec les différentes versions du système Divergence.

Il est très interessant de noter que les meilleurs résultats sont obtenus pour le système Divergence2 et qu'il correspond à un étiquetage en parole continue.

A titre indicatif, le taux de reconnaissance monolocuteur sur un ensemble test différent de l'ensemble d'apprentissage est de 88% ; ce faible taux est en grande partie une des explications de nos performances moyennes. Il est certain que notre ensemble d'apprentissage doit être très largement étoffé pour permettre d'avoir un dictionnaire de meilleure qualité, ce que nous faisons actuellement.
Malgrè tout, il ne s'agit que de premières expérimentations, et il faut espérer que les résultats puissent être améliorés.

4.2. SYSTEME SALAM

Le pré-traitement acoustique (système SED) ne nécessite aucun apprentissage (indépendant des locuteurs, des langues, et des corpus). Pour nos expériences, nous disposions des corpus français d'EUROM0 (23 phrases prononcées en continu par deux locuteurs) et de BDSONS (chiffres isolés prononcés 10 fois par dix locuteurs). Les phrases prononcées par un locuteur d'EUROM0 et les chiffres prononcés parles dix locuteurs de BDSONS ont été utilisés pour l'apprentissage des MMC élémentaires, le deuxième locuteur d'EUROM0 a été utilisé en phase de test (locuteur inconnu pour le système).

Les expériences menées se portent successivement sur l'utilisation de 23 classes phonétiques (**SALAM2-23**), la division des segments de 100 ms en trois sous-segments de longueur identiques (**SALAM2-23-a**), l'adaptation au locuteur sans division des segments (**SALAM2-23-b**), l'adaptation au locuteur et la division des segments (**SALAM23-ab**). Les meilleurs résultats sont obtenus avec la combinaison de l'adaptation au locuteur et la division des segments où nous obtenons 78% d'alignment correct à ±20 ms. Il faut noter que de plus le taux d'erreur grave a sensiblement baissé : nous sommes passés de 8% à 5% (Cf. tableau 5). A titre indicatif nous avons complété ce tableau avec les résultats trouvés antérieurement à l'aide de MMC discrets (**SALAM1**).

	étiquetage correct ±20 ms (%)	erreurs graves ±50 ms (%)
SALAM1 discrets	76	8
SALAM1 lissés	79	6
SALAM2-23	75	8
SALAM2-23-a division segment	76	7
SALAM2-23-b adaptation-loc	77	6
SALAM2-23-ab	78	5

Tableau 5 : Résultats obtenus avec les différentes versions du système SALAM.

5. COMMENTAIRES ET CONCLUSION

Etant donné la nature des expériences que nous avons menées, il est très difficile voire impossible de comparer ces deux systèmes.

Les premiers résultats obtenus avec le système DIVERGENCE sont encourageants.
Une optique de travail interessante est d'utiliser une modélisation Markovienne continue, nous permettant ainsi d'éliminer l'étape de quantification vectorielle qui parait induire des erreurs ; et de travailler ainsi directement sur la paramétrisation cepstrale des segments. De plus, nous introduirons la notion de durée d'un son [Suaudeau-93] ce qui devrait être un apport considérable.

Les résultats obtenus avec le système SALAM confirment sa robustesse et ses bonnes performances pour les tests indépendants du locuteur. Les deux améliorations proposées (division des segments trop longs et adaptation non supervisée au locuteur) paraissent très prometteuses pour l'avenir. Nous étudions actuellement une approche plus élaborée de la procédure de division des segments longs qui nous permettra d'augmenter les performances du système.

REFERENCES BIBLIOGRAPHIQUES

[André-Obrecht-88] R. André-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE trans. on Acoustics Speech & Signal Processing, vol. 36, n° 1, pp. 29-40, January 1988.

[André-Obrecht-92] R. André-Obrecht, G. Pérennou, N. Vigouroux, "Deux approches de l'étiquetage en événements phonétiques", 19ème Journées d'Etude sur la Parole, pp. 279-284, Bruxelles, mai 1992.

[Boë-88] L.J. Boë, L. Miclet, C. Sorin, "L'étiquetage de la base de données des sons du français du greco commnucation parlée", Premières journées nationales du GRECO-PRC Communication Parlée, pp. 39-47, Paris, Novembre 1988.

[Farhat-92] A. Farhat, G. Pérennou, N. Vigouroux, "Segmentation en événements phonétiques et modèles Markoviens HMM pour l'étiquetage phonotypique", 19ème Journées d'Etudes sur la Parole, pp. 297-302, Bruxelles, mai 1992.

[Farhat-93] A. Farhat, G. Pérennou, R. André-Obrecht, "A Segmental Approach Versus a Centisecond One For Automatic Phonetic Time-Alignment", 3th European Conference on Speech Communication and Technology, pp. 657-660, Berlin, septembre 1993.

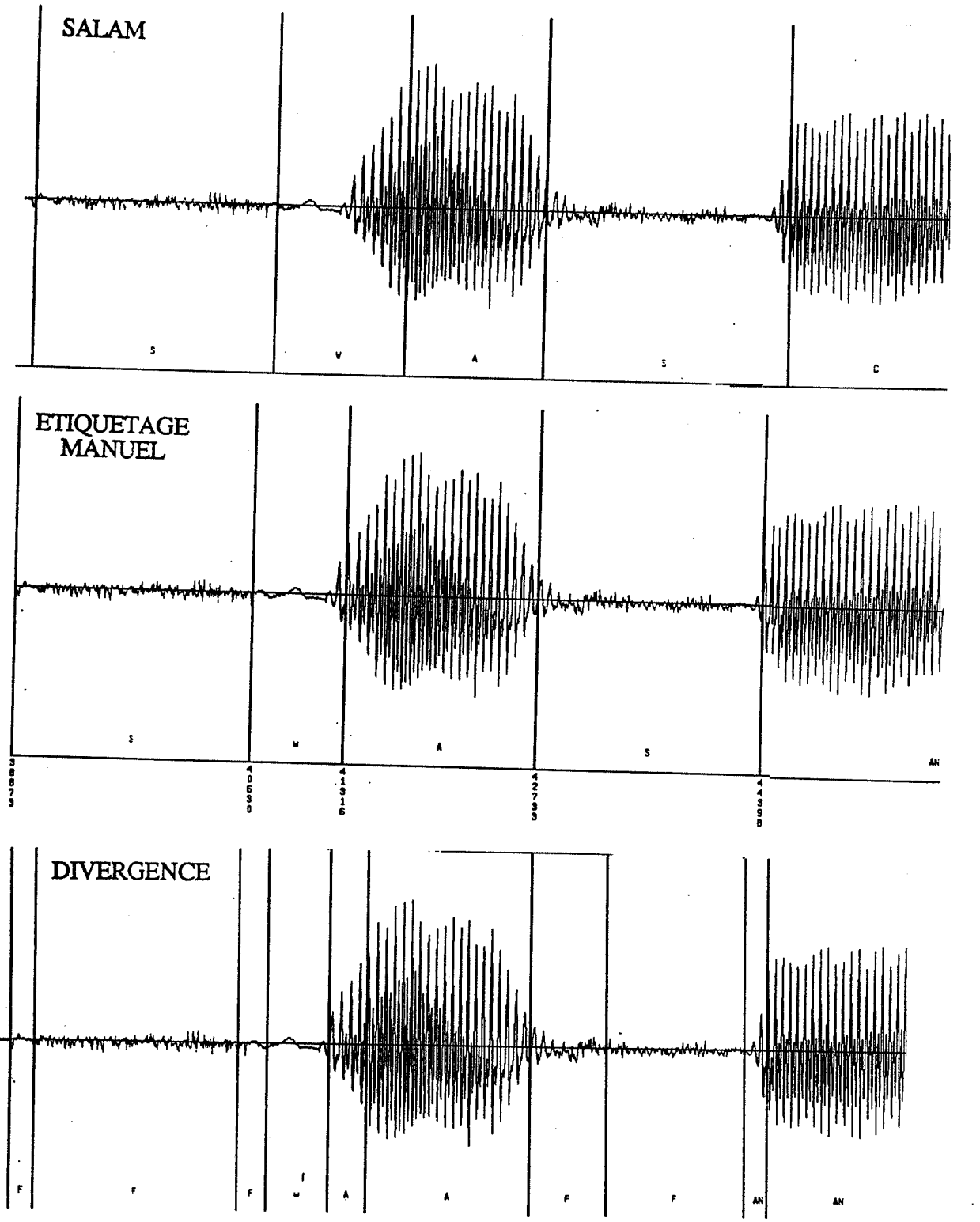
[Leung 84] H.C Leung, V.W. Zue, "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", Proc. of the International Conference on Acoustic Speech and Signal Processing, pp. 2.7.1-2.7.4, San-Diego, 1984.

[Pérennou-91] G.Pérennou, H. Kabré, N. Vigouroux, "Automatic SAPHO Segmentation of EUROM0, Multi-Lingual Speech Corpora Into Phonetics Events, SAM-Report, June 1991.

[Su-87] H.Y. Su, "Reconnaissance acoustico-phonétique en parole continue par quantification vectorielle ; adaptation du dictionnaire au locuteur", Thèse de l'université de Rennes I, novembre 1987.

[Suaudeau-93] N.Suaudeau, R.André-Obrecht, "Sound duration Modelling and Time-Variable Speaking rate in a Speech Recognition System", 3rd European Conference on Speech Communication and Technology, pp. 309-311, Berlin, septembre 1993.

Exemple d'étiquetage avec les deux systèmes SALAM et DIVERGENCE en comparaison avec l'étiquetage de l'expert



COMPARAISON DE DEUX METHODES NON PARAMETRIQUES POUR LA DETECTION DES OCCLUSIVES SOURDES

F. MALBOS*, R. ANDRE-OBRECHT**, M. BAUDRY*

* LIUM, UNIVERSITE DU MAINE,
AVENUE OLIVIER MESSIAEN, 72017 LE MANS
** IRIT, UNIVERSITE PAUL SABATIER,
118 ROUTE DE NARBONNE, 31062 TOULOUSE

Résumé

Le travail présenté s'insère dans le cadre très large du décodage acoustico-phonétique en parole continue. L'idée principale est d'exploiter au niveau du prétraitement, les techniques issues du traitement de signal afin de détecter des événements robustes et fiables qui guideront les étapes ultérieures de reconnaissance.

Deux méthodes non paramétriques sont proposées afin de détecter les occlusives sourdes dans le discours continu et plus précisément leur barre d'explosion. La première utilise la transformation en ondelettes (ondelette de Morlet). La deuxième méthode utilise la transformée de Fourier pour calculer deux indices fréquentiels.

Ces deux approches ont été testées sur un corpus issu de BDSONS. Les taux d'omissions et de fausses détections permettent de les évaluer. Il apparaît que pour éviter tout risque d'omissions, il reste 10 à 15 % de détections correspondant à de fausses détections (typiquement les débuts du son */s/*).

Mots-clés : *détection, occlusives, ondelette, Fourier*

INTRODUCTION

Dans un système de reconnaissance automatique de parole, le décodage acoustico-phonétique (DAP) est une des étapes clés : la performance globale du système dépend de manière souvent dramatique de la performance de ce module. Comme son nom l'indique, le DAP consiste à transcrire le signal acoustique en une suite discrète de symboles ou d'événements phonétiques. Un de ses premiers objectifs est d'extraire des informations robustes afin de les utiliser comme des îlots de confiance pour la suite du traitement d'identification. L'examen visuel de sonagrammes nous laissent penser que les occlusives sourdes doivent faire partie de ces points d'ancrage [Caelen &al 88], [François 92] et que, de part leur évolution spectrale, les approches de type *détection de rupture fréquentielle* doivent être appropriées.

Deux détecteurs sont développés dans ce sens, l'un est basé sur la transformation en ondelettes et l'autre sur la transformation de Fourier classique. La comparaison de ces approches et leurs résultats confirment que les méthodes de traitement de signal couplées à des connaissances de nature phonétique peuvent se révéler très efficaces.

ANALYSE EN ONDELETTES

PRINCIPE DE LA METHODE

Présentation de la méthode

Afin de simplifier la détection de la barre d'explosion de l'occlusive, cette impulsion glottique est modélisée par un bruit impulsionnel. Parmi les représentations temps-fréquence, la transformée en ondelettes a été choisie en raison de la minimisation de l'incertitude temporelle qu'elle offre en hautes fréquences.

L'ondelette continue

La transformée en ondelettes associe à un signal réel p(t) une fonction O(b,a) complexe et bidimensionnelle dont les arguments b et a (a>0) représentent respectivement le temps et le facteur d'échelle. Les coefficients O(b,a) de la transformée en ondelettes sont calculés de la façon suivante :

$$O(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} h^*\left(\frac{t-b}{a}\right) p(t)dt \quad (1)$$

avec :

h(t) : ondelette mère,
$h^*(t)$: complexe conjugué de h(t).

En notant, $\langle , \rangle$, le produit scalaire pour les signaux continus, (1) s'écrit :

$$O_{b,a} = \langle p, h_{b,a} \rangle \quad (2)$$

En posant, $h^-(t) = h^*(-t)$ et * l'opérateur de convolution, (1) s'écrit :

$$O_{b,a} = p * h^-_{a,b} \quad (3)$$

Les expressions (2) et (3) permettent d'interpréter la transformée en ondelettes de la façon suivante :
– une décomposition du signal sur une famille de fonctions se déduisant les unes des autres par contraction et dilatation de l'ondelette "mère",
– un banc de filtres linéaires dont chaque cellule présente une résolution fréquentielle relative constante.

Calcul des coefficients d'ondelettes

Parmi les nombreuses ondelettes utilisées en traitement du signal, celle de Morlet minimise la relation d'incertitude temps-fréquence offrant ainsi le meilleur compromis entre les localisations temporelle et fréquentielle [Dorize 89]. L'ondelette de Morlet notée h_m est définie de la façon suivante :

$$h_m(t) = e^{-t^2/2} e^{i\omega_o t} \quad (4) \quad \text{avec } \omega_o = 5$$

Dans notre application, les coefficients sont calculés à l'aide d'un algorithme rapide utilisant un filtre à réponse impulsionnelle infinie [Barrat 92]. L'ondelette de Morlet est approximée par une fonction h(t) de la forme :

$$h(t) = (1+\sigma|t|)e^{-\sigma|t|}e^{ict} \quad (5) \quad \text{avec } c = 5 \text{ et } \xi = a^{-1}$$

En notant T la période d'échantillonnage, les coefficients d'ondelette sont définis par :

$$O(kT,a) = \frac{T}{a}\sum_n p(nT)h^*\left[\frac{(n-k)T}{a}\right] \quad (6)$$

Afin de simplifier leur écriture, ils sont notés $O_i(n)$. Le paramètre i désigne le numéro de la voie d'analyse.

Le module de la transformée en ondelettes (scalogramme) d'une occlusive sourde est présenté à la figure 1.

Etude des fonctions de corrélation

Notre système de détection est basé sur l'étude des fonctions de corrélation normalisées notées C_i entre les modules de chaque réponse impulsionnelle h_i du banc de filtres et les modules de la transformées en ondelettes O_i du signal de parole. La normalisation énergétique de la fonction C_i a pour objet de s'affranchir de la différence d'amplitude entre les phonèmes de forte énergie (voyelle) et les occlusives.

$$C_i(p) = \frac{\sum_{n=0}^{n=N_o} |O_i(n)||h_i(n+p)|}{\sqrt{\sum_{n=0}^{n=No} O_i^2(n)}\sqrt{\sum_{n=0}^{n=No} h_i^2(n)}} \quad (7)$$

Chaque réponse impulsionnelle du banc de filtres étant décrite par une fonction du type exponentielle décroissante, son module peut être approximé par zéro pour tous les instants t supérieurs à une valeur limite t_{li}. Cette approximation est capitale et explique la présence pour chaque fonction Ci d'un minimum local m_{ki} précédant un maximum local M_{ki} dont l'amplitude est proche de l'unité (figure 2).

Ce phénomène apparaissant dans toutes les voies, il est intéressant d'utiliser la notion de banc de filtres pour construire deux signaux S_M et S_m tels que :
– S_M est la somme temporelle synchrone de toutes les fonctions C_i. Ses maxima sont synchrones avec le bruit impulsionnel recherché dans le signal de parole.

$$S_M(n) = \sum_{i=1}^{i=16} C_i(n) \quad (8)$$

– S_m est la somme asynchrone suivant une loi temporelle théorique inversement proportionnelle au facteur d'échelle a (n_i=f(i)).

$$S_m(n) = \sum_{i=1}^{i=16} C_i(n-n_i) \quad (9)$$

avec : $n_i > 0$ pour $i \neq 16$ et $n_i = 0$ pour $i = 16$

MISE EN OEUVRE

Il est primordial de visualiser les signaux S_M et S_m dans le cas d'un signal de parole (figure 3). Leur visualisation pour un nombre important d'occlusives fait apparaître les propriétés suivantes :
- la présence pour le signal S_m d'un minimum local de très faible amplitude précédant la barre d'explosion de l'occlusive,
– l'existence pour le signal S_M d'un maximum local synchrone avec le "burst" de l'occlusive,
– une détection visuelle plus aisée des minima de la courbe S_m par rapport aux maxima de la courbe S_M,

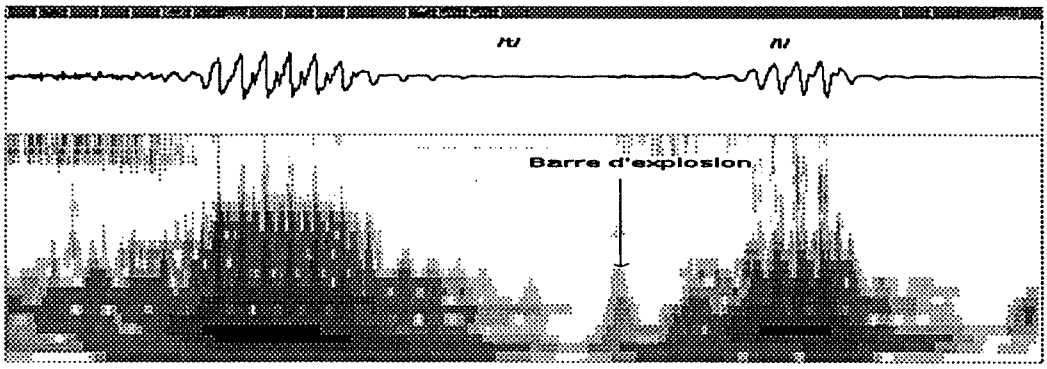


Figure 1

Signal de parole

Scalogramme d'une occlusive sourde

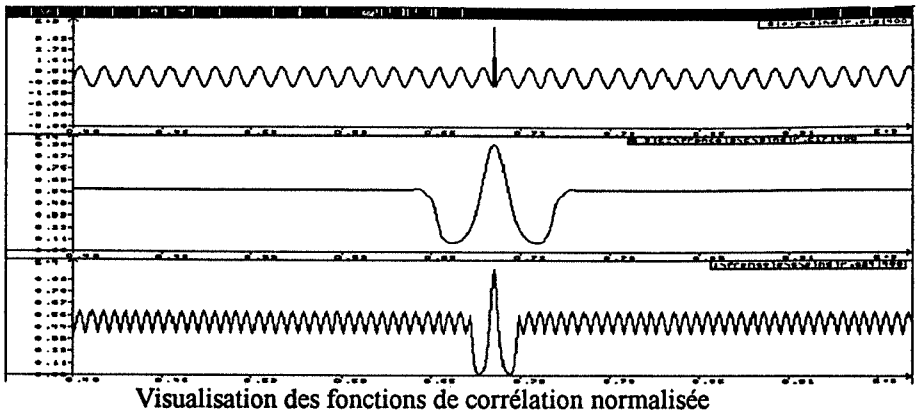


Figure 2a: Insertion d'un pic de Dirac dans un sinus de fréquence 625 Hz

Figure 2b : Fonction de corrélation normalisée pour une fréquence de 1487 Hz

Figure 2c : Fonction de corrélation normalisée pour une fréquence de 5000 Hz

Visualisation des fonctions de corrélation normalisée
Figure 2

– une meilleure pertinence de la première propriété par rapport à la seconde.

L'étude de ces courbes nous a permis de concevoir notre algorithme de détection. Il est composé de deux modules :

– la recherche sur le signal S_m d'un minimum local m d'amplitude inférieure à une valeur limite S,
– la recherche au sein du signal S_M d'un maximum local M postérieur au minimum m. Cette démarche permet une bonne localisation temporelle de la barre d'explosion.

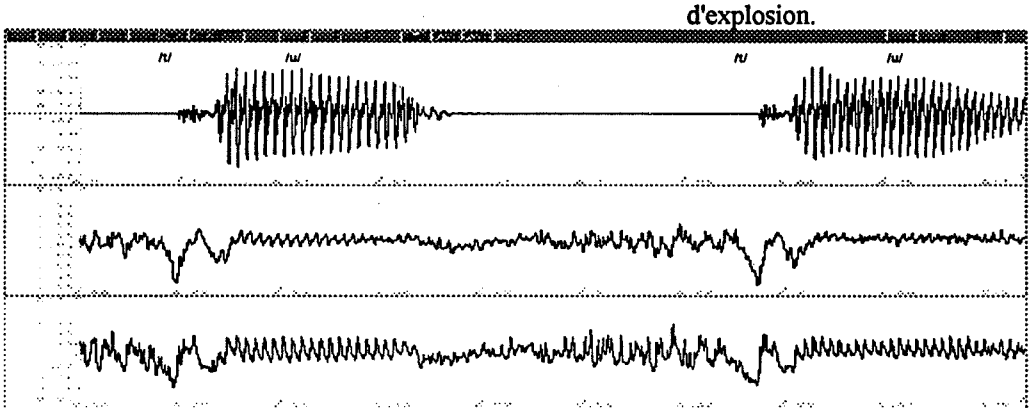


Figure 3a
Signal de parole

Figure 3b
Signal S_m

Figure 3c
Signal S_M

Fonctions S_m et S_M pour un signal de parole
Figure 3

Décomposition fréquentielle

La mise au point de la méthode de détection des occlusives sourdes a été effectuée à l'aide d'un corpus de mots enchaînés échantillonnés à 10 kHz [Malbos &al 94]. Le corpus présenté dans cet article possédant une fréquence d'échantillonnage de 16 kHz, un sous échantillonnage à 10 kHz a été accompli [Crochiere 89]. Le choix d'une décomposition des voies d'analyse en quart d'octave sur quatre octaves [372 Hz, 5000 Hz] se justifie de la façon suivante :
– une résolution fréquentielle relative de l'ondelette "mère" située entre le tiers et le quart d'octave,
– une concentration énergétique (Halle [57]) :
dans la bande 500 Hz et au dessus de 4000 Hz pour les dentales,
en moyennes fréquences [1500-4000] Hz pour les vélaires,
dans la bande [500-1500] Hz pour les labiales.

Prétraitement

Le corpus comprend de temps à autre des zones de silence d'énergie nulle entre deux mots isolés. Cela engendre l'apparition de minima artificiels au sein de la fonction S_m. Afin de ne pas les confondre avec ceux correspondant aux occlusives, ces zones de silence ont été remplacées par un bruit aléatoire. Son énergie est voisine de celle d'un silence occlusif. La neutralité de ce bruit par rapport à notre détecteur a été testée. Il n'est en aucun cas responsable de fausses détections.

Plusieurs critères énergétiques ont été implémentés afin de délimiter les zones du signal de parole susceptibles de contenir une occlusive sourde. Parmi ceux-ci, nous pouvons citer :
– la recherche automatique edu début et de la fin de la phrase,
– la localisation de différentes zones Z d'une durée de 15 ms susceptibles de contenir une partie du silence des occlusives sourdes et leur barre d'explosion.

La première étape est la recherche dans les zones Z de tous les minima (au sens de la dérivée) d'amplitude inférieure ou égale à une valeur limite S. Un "critère de distance" a été implémenté. Son rôle est double :
– conserver un minimum parmi les minima correspondant à une même occurrence,

– supprimer les détections d_j postérieures à une détection d_i si la l'écart temporel entre ces différents événements est inférieure à la distance minimale entre les barres d'explosion de deux occlusives sourdes consécutives.

L'étude des signaux représentant les mots "aptitude" et "activité" prononcés rapidement par plusieurs locuteurs a montré que la valeur minimale entre deux barres d'explosion d'occlusives est approximativement de 140 ms. Cette évaluation ayant porté sur un vocabulaire réduit et un nombre restreint de locuteurs, la valeur limite utilisée dans notre système de détection a été fixée à 100 ms.

D'autres critères ayant pour but de diminuer les fausses détections ont été implémentés mais n'ont pas été retenus. Nous pouvons citer celui du "minimum minimorum". Soit deux minima voisins Ai et Aj de la courbe S_m (Ai est antérieur à Aj). Il peut être logique d'associer à une occlusive éventuelle celui d'amplitude la plus faible. Ce raisonnement est souvent mis en défaut dans le cas d'une occurrence "occlusive-voyelle". Le premier cycle de voisement de la voyelle peut en effet présenter au sein de la courbe S_m un minimum plus marqué que celui correspondant à l'occlusive.

ANALYSE PAR TRANSFORMEE DE FOURIER

Cette méthode a été développée suite aux résultats obtenus par un algorithme de segmentation automatique du signal [André-Obrecht 88]. A l'aide de cette méthode, les plosives sourdes sont décomposées systématiquement en trois segments:

– une zone transitoire entre le son précédent et le silence de la plosive, qui correspond à la fermeture des cordes vocales,

– la zone de silence,

– un segment dont les frontières coïncident avec l'explosion de la consonne et le début du son suivant.

Des observations semblables sont faites à propos des plosives sonores mettant en évidence la barre de voisement et la barre d'explosion lorsque celle-ci existe.

Nous avons naturellement cherché à extraire de l'ensemble des frontières détectées automatiquement celles qui étaient synonymes d'une explosion pour ensuite localiser la plosive dans son ensemble.

La méthode décrite ci-dessous s'applique pour la détection des plosives sourdes et sonores.

PRINCIPE DE LA METHODE

L'algorithme de détection est basé sur le couplage de trois modules :

– la segmentation automatique du signal de parole,

– le calcul d'un indice de voisement sur chaque segment,

– le calcul de deux indices fréquentiels suivi d'un filtrage et seuillage.

La description du module de segmentation automatique comme la description du calcul de voisement ne seront pas décrits puisqu'ils ont fait l'objet de publications antérieures [André-Obrecht 88], [Deléglise 91].

Les indices fréquentiels

L'objectif est de localiser la barre d'explosion d'une plosive, c'est à dire un accroissement d'énergie haute fréquence suivi d'un relâchement, après avoir détecté une zone d'énergie faible dans la bande formantique (500Hz-3000Hz).

Après l'application d'une préemphase sur le signal et d'une fenêtre de Hamming, une transformée de Fourier est effectuée sur chaque trame *n* de parole (fenêtre d'environ 20ms). L'énergie dans 24 bandes de fréquence est calculée ; ces canaux sont répartis sur l'axe fréquentiel selon l'échelle Mel. Deux indices en sont extraits :

– l'indice formantique f_n est une somme des énergies véhiculées dans la bande formantique élargie, à savoir dans les canaux compris entre 500Hz et 4000Hz.

– après calcul de la variable $d_n$correspondant à la variation d'énergie entre deux trames consécutives pour les fréquences supérieures à 4000Hz, un filtrage temporel est appliqué pour finalement obtenir l'indice delta Δ_n, selon la formule suivante :

$$\Delta_n = \alpha\Delta_{n-1} + \beta\Delta_{n-2} + d_n \quad (10)$$

MISE EN OEUVRE

L'algorithme se détection se décompose en deux phases.

Détection d'éventuelles explosions.

Le calcul des indices et la segmentation du signal de parole sont effectués en parallèle. Dès qu'est détectée une zone de faible énergie formantique ($f_n < \lambda 1$) correspondant au moins à un segment, et aussitôt que ce même seuil est franchi, est recherché le prochain maximum de l'indice delta. Si ce maximum dépasse un seuil $\lambda 2$, s'il est suffisamment proche d'une frontière et si la décroissance de delta est ensuite suffisamment rapide, la frontière la plus proche du maximum est proposée comme début de l'explosion d'une plosive.

Validation de la détection.

Afin de valider ou non une détection d'explosion, chaque segment est identifié comme voisé, non voisé ou silence par un test de voisement.

Sont validées :

- comme explosion de plosives sourdes toute frontière qui suit un segment de silence et qui précède un segment bruité,
- comme explosion de plosives sonores toute frontière située entre deux zones voisées, la zone voisée la précédant devant être de longueur supérieure à 40ms.

EVALUATION

CORPUS

Le corpus est un sous-ensemble des logatomes enregistrés dans BDSONS. Ses caractéristiques sont :
- 522 mots isolés de type CVCV (consonne voyelle consonne voyelle)
- une élocution multilocuteurs (1 homme, 3 femmes),
- un contexte /a/, /i/, /u/,
- une fréquence d'échantillonnage de 16 kHz,

La répartition des occlusives sourdes, sonores et des autres consonnes est présentée dans les tableaux 1 et 2 :

Occlusives sourdes		Occlusives sonores	
p	62	b	52
k	74	g	50
t	98	d	70
total	234	total	172

tableau 1

Autres consonnes			
m	48	s	32
n	30	S	48
N	34	v	42
l	66	z	50
r	54	Z	56
f	0	total	460

Tableau 2

Afin d'évaluer les deux détecteurs, nous avons déterminé pour chacun d'entre eux :
- le nombre d'événements non détectés (omissions),
- le nombre de fausses détections

RESULTATS DE LA METHODE TEMPS ECHELLE

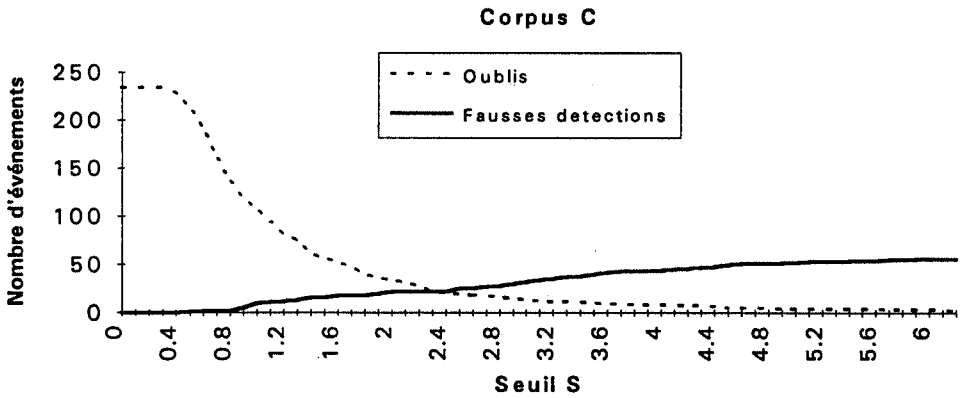
La notion de taux d'erreur égal (nombre d'oublis égal au nombre de fausses détections) est une notion intéressante pour l'évaluation d'un détecteur d'événements. Elle a été utilisée pour la validation d'un détecteur de bruits impulsionnels dans le cadre des enregistrements anciens (Montresor [91]). Les caractéristiques de notre détecteur sont présentées à l'aide de la figure 4. Le taux d'erreur égal correspond à un seuil S=2,3 et un taux de détection égal à 90,6%. Les fausses détections correspondent principalement aux phonèmes /s/, /R/, /z/. Une évaluation de notre détecteur a été réalisée sur un ensemble de phrases phonétiquement équilibrées. Le taux d'erreur égal pour les occlusives sourdes correspond à un taux de détection de 83,2% et à un seuil S=3,2.

RESULTATS DE LA METHODE TEMPS FREQUENCE

Une phase d'apprentissage a été nécessaire pour fixer les deux seuils $\lambda 1$ et $\lambda 2$ ainsi que les paramètres du filtre α et β Les données d'apprentissage sont totalement indépendantes du corpus de test ; elles correspondent à dix phrases phonétiquement équilibrées enregistrés à une fréquence de 12,8 kHz et prononcées par un homme.

Le système de détection a ensuite été expérimenté sur le corpus décrit précédemment. Il en résulte que la phase de validation a permis d'éliminer 22% des détections sans provoquer d'omissions supplémentaires. 100% des plosives sourdes sont correctement localisées.

Sur les fausses détections qui représentent un taux de 14% par rapport au nombre global de détections retenues, un tiers est identifié comme des plosives sourdes, ce qui représente un taux de 12% de fausses détections pour cette catégorie de sons. Ces erreurs correspondent à des débuts de mots commençant par les sons /R/, /s/ et pour la plupart /ch/ (nombre d'entre elles sont peut être dues au procédé de concaténation de signal utilise dans BDSONS).



Caractéristiques du détecteur "temps-échelle"
Figure 4

A titre indicatif, pour les plosives sonores, les premiers résultats donnent un taux de détection de 20%.

Dans la mesure où le corpus testé ne comprend que des mots isolés, enregistrés dans de très bonnes conditions, de nouvelles expériences doivent être faites sur de la parole continue, dans des conditions autres.

Une première expérience a été réalisée sur des phrases issues du corpus EUROM0. les résultats précédents semblent être confirmés pour les plosives sourdes : un taux de fausses détections de l'ordre de 14% est observé pour un taux d'omissions de 2,6%.

CONCLUSION

Les deux détecteurs que nous avons mis en oeuvre reposent sur des méthodes non paramétriques.

La première méthode utilise la transformée en ondelettes. Basée sur une représentation temps-échelle, elle met en évidence le fait que la barre d'explosion des occlusives peut être modélisée par un bruit impulsionnel.

La seconde méthode utilise la transformée de Fourier basée sur une représentation temps-fréquence. Construite à partir de résultats de segmentation de parole, elle utilise différents paramètres énergétiques.

Les détecteurs offrent deux caractéristiques principales : d'une part un bon compromis alliant un fort taux de détection à un petit nombre d'omissions, et d'autre part une localisation très précise de la barre d'explosion avec une précision inférieure à 1 ms.

Au vu de ces résultats plutôt encourageants, il serait particulièrement intéressant d'étendre ce travail dans le cadre de la détection des occlusives sonores.

REFERENCES BIBLIOGRAPHIQUES

[André-Obrecht 88] : ANDRE-OBRECHT., "A new statistical approach for the automatic segmentation of continous speech signals", IEEE Trans. on Acoustic, Speech, Signal Processing, vol 36, n°1, Janvier 1988.

[Barrat 92] :BARRAT M., LEPETIT O. : "Calcul rapide de la transformée en ondelettes", Traitement du signal ,1992, 8, 1, pp.43-49.

[Caelen 88] CAELEN J.& AL : "Une base de règles pour le décodage acoustico-phonétique : le cas des occlusives sourdes", Actes du séminaire GRECO-PRC - Université de Nancy, 1988.

[Crohiere 89] : CROCHIERE R., RABINER L. : "Multirate Digital Signal Processing", Prentice-hall signal processing series, Oppenheim Alan V. editor, 1983.

[Deléglise 91] : DELEGLISE P: "Une architecture logicielle pour le décodage acoustico-phonétique, application à la détection d'événements phonétiques" Thèse de doctorat d'état, Université Paris VI, 1991.

[Dorize 89]DORIZE C., GRAM-HANSEN K. : "Related positive time-frequency energy distributions", Wavelets and applications, Editor Y Meyer, Masson Spinger verlag, 1989.

[Dutilleux 89] : DUTILLEUX P. "An implementation of the "algorithme à trous" to compute the wavelet transform", Wavelets, time frequency Methods and Phase Space? Springer Verlag, 1989.

[François 92] FRANCOIS D., FOHR D., "Contribution de réseaux neuronaux pour la reconnaissance des occlusives au sein du système expert Aphodex", 19ème J.E.P (Bruxelles)., pp. 405-408, 1992.

[Halle 57] : HALLE M., HUGUES G.W., RADLEY J.-P.A. : "Acoustic Properties of stop consonants" J. Acoust. Soc. Am. 29, 1, pp. 107-116.

[Holschneider 87] : HOLSCHNEIDER M., KRONLAND-MARTINET R., MORLET J., TCHAMITCHIAN Ph : " A real-time algorithm for signal analysis with the help of the wavelet transform", Wavelets, time frequency methods and Phase Space, Springer-Verlag, 1989.

[Malbos &al 94] : MALBOS F., BAUDRY M., MONTRESOR S. : "Détection des occlusives à l'aide de la transformée en ondelettes", 3ème congrès de la S.F.A, Toulouse, 1994 (à paraitre).

[Montresor 90] : MONTRESOR M., VALIERE J.C, BAUDRY M. : "Détection et suppression de bruits impulsionnels appliqués à la restauration d'enregistrements anciens", 1er congrès de la S.F.A., Lyon, 1990.

CHOIX ET ADAPTATION D'UN PHONETISEUR POUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

K. BARTKOVA, D. LARREUR, I. METAYER

FRANCE TELECOM, CNET, LANNION, LAA/TSS/RCP
ROUTE DE TREGASTEL, 22300 LANNION, FRANCE

Résumé

Dans cette étude nous comparons trois phonétiseurs en vue de choisir celui le mieux adapté pour la reconnaissance de la parole. Ces trois systèmes de phonétisation sont: le dictionnaire BDLEX, le dictionnaire SLC et un phonétiseur par règles dérivé du système de synthèse du CNET. Les évaluations se sont déroulées dans un premier temps au niveau théorique par une comparaison de la transcription manuelle avec les transcriptions automatiques. Dans un deuxième temps, nous avons comparé les performances d'un système de reconnaissance de la parole en utilisant les trois transcriptions automatiques avec les performances obtenues par la transcription manuelle. Les tests des transcriptions nous ont permis de cibler nos besoins et d'entreprendre l'adaptation du système choisi pour répondre aux besoins de la reconnaissance de la parole.

Mots-clés: *système de transcription phonétique, reconnaissance automatique de la parole.*

INTRODUCTION

Afin de créer des réseaux d'unités plus petites que le mot, on est amené en reconnaissance de la parole à transcrire phonétiquement chaque unité lexicale en terme d'unités phonologiques, le plus couramment en phonèmes. Cette procédure de transcription s'est effectuée chez nous, jusque-là, à la main. Une telle démarche nécessite des connaissances phonétiques non négligeables et même si l'on dispose de telles connaissances, on peut facilement introduire des erreurs dues à l'inattention ou à la fatigue car la transcription phonétique des bases de données est souvent longue et fastidieuse. Afin d'éviter de telles erreurs et d'accélérer la transcription phonétique, nous avons décidé d'utiliser une transcription phonétique automatique fournie soit par un dictionnaire phonétique soit par un système de phonétisation par règles.

Pour cibler nos besoins en reconnaissance de la parole (qui sont sensiblement différents de ceux de la synthèse de la parole) et voir comment ceux-ci pouvaient être traités, nous avons testé 2 dictionnaires phonétiques et un système de phonétisation par règles. Ces trois systèmes étaient les suivants:

- ***phonétiseur par règles** (dérivé du phonétiseur utilisé pour la synthèse de la parole au CNET)[1];*
- ***dictionnaire SLC** (dictionnaire utilisé pour le traitement de la langue naturelle au CNET)[2];*
- ***dictionnaire BDLEX** (du GDR-PRC "Communication Homme-Machine")[3].*

Les deux dictionnaires comportaient une liste de mots phonétisés ainsi que quelques renseignements linguistiques sur le mot mais pas de règles de phonétisation. L'inconvénient des dictionnaires est qu'en cas d'absence d'un mot dans leur lexique, le mot n'est pas phonétisé du tout. Le système de phonétisation par règles comportait un lexique d'exceptions (pour les mots qui constituent une exception de par leur prononciation ou qui se verraient attribuer une catégorie grammaticale fausse de par leur terminaison) et des règles de phonétisation pour le français.

PROCEDURES D'EVALUATION

La qualité de la phonétisation des mots, fournie par les trois systèmes de phonétisation, a été évaluée au niveau de la comparaison avec la transcription manuelle (comparaison "théorique" ou "qualitative") et au niveau de la performance du système de reconnaissance (comparaison "pratique" ou "quantitative"). En tout, nous avons évalué la performance des trois phonétiseurs sur trois niveaux :

- ***au niveau de la couverture des corpus phonétisés ;***
- ***au niveau de la "ressemblance" de la transcription avec la transcription manuelle ;***
- ***au niveau des performances fournies par le système de reconnaissance de la parole à base des modèles de Markov (HMM) utilisant les transcriptions automatiques.***

Les corpus à phonétiser, utilisés lors des différents tests, et le nombre de locuteurs qui ont enregistré les bases de données correspondantes étaient les suivants:

- **allophones** *(680 phrases courtes, 1000 locuteurs)[4] ;*
- **500 mots isolés** *(les plus courants du français, 10 locuteurs, et 3 répétitions par locuteur);*
- **Trégor** *(36 mots isolés et expressions, 765 locuteurs);*

- **chiffres** *(10 mots isolés, 770 locuteurs);*
- **nombres à deux chiffres** *(de 00 à 99, 789 locuteurs);*
- **trains** *(phrases, vocabulaire de 252 mots, 120 locuteurs).*

Pour évaluer le taux de couverture des phonétiseurs et la ressemblance de la phonétisation automatique avec la phonétisation manuelle, nous avons également employé le corpus **"RCP"** qui contient plus de 1000 mots et expressions utilisés pour le développement des services vocaux.

COUVERTURE DES CORPUS FOURNIE PAR LES PHONETISEURS

Le taux de couverture fourni par le phonétiseur par règles était de 100% pour les différents corpus. Les taux de couverture des deux autre phonétiseurs (dictionnaire BDLEX et SLC) sont représentés sur la figure ci-dessous.

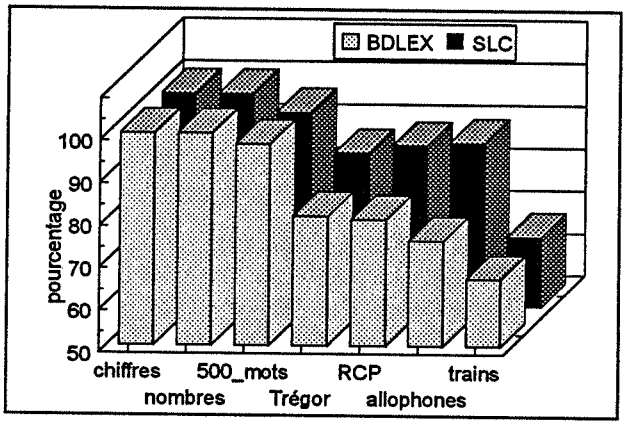


Figure 1: *Taux de couverture des dictionnaires BDLEX et SLC sur les différents corpus.*

La majorité de mots non phonétisés des corpus étaient des noms propres: les dictionnaires SLC et BDLEX ne traitent ni les sigles ni les noms propres. Mais du lexique de BDLEX manquait un nombre considérable de mots courants comme par exemple : **"sondage", "opérateur", "serveur", "réservation", "enregistreur", "identifier", "répondeur"...**

COMPARAISON "THEORIQUE" DES PHONETISATIONS MANUELLE ET AUTOMATIQUE

Procédure de comparaison

La comparaison "théorique" entre la phonétisation manuelle et la phonétisation automatique a été faite uniquement sur le corpus RCP. Chaque phonétisation automatique a été comparée par programmation dynamique à la phonétisation manuelle, fournie par une phonéticienne. La transcription manuelle correspond à peu près à ce que l'on trouve dans [5] avec des variantes de prononciation, reflétant une articulation plus soignée et plus coarticulée ([6, 7, 8]). L'écart entre la phonétisation manuelle et celle fournie par le système automatique a été exprimée par le nombre de substitutions, d'insertions et d'omissions de phonèmes.

Lors de l'alignement de deux phonétisations, la pénalisation était moins forte pour la substitution d'une voyelle par une autre ayant une aperture différente ([e] avec [ɛ], [o] avec [ɔ]...) que pour les autres sortes de substitutions afin de favoriser l'alignement de ces deux voyelles. La raison pour cela est la suivante: en français (même du bon usage), à l'intérieur des mots (position supposée non accentuée) la réalisation des deux apertures n'est pas très nette et il y a une tendance vers une centralisation surtout quand la durée vocalique tend vers une durée "courte" (voyelle fermée moins fermée, voyelle ouverte moins ouverte). Par ailleurs, les deux timbres n'assurent plus de distinction phonologique en français moderne. Ainsi leur substitution mutuelle doit être moins pénalisée que la substitution de deux phonèmes appartenant à des catégories de phonèmes plus hétérogènes.

Nous avons pénalisé très peu les insertions de la "présence facultative" (de la latence) d'un phonème. Par contre l'omission de la latence a été fortement pénalisée. La raison pour cela est la suivante: si un phonème est donné latent par le phonétiseur, le réseau contenant la description phonologique du mot dans le système de reconnaissance automatique de la parole contiendra deux branchements à cet endroit, un autorisant le passage par ce phonème et donc sa modélisation acoustique et l'autre évitant ce même phonème ainsi que sa modélisation (voir Figure 2). Par contre si cette propriété de latence n'est pas indiquée par le phonétiseur, la passage par le réseau acoustique modélisant le phonème devient obligatoire. Cela cause en général un décalage au niveau de la segmentation du signal ainsi qu'au niveau de l'alignement temporel du mot à reconnaître avec le modèle de référence et cela peut s'avérer fatal pour la bonne reconnaissance.

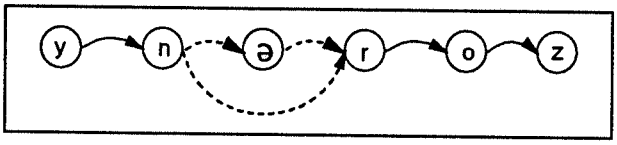


Figure 2: *Début du réseau phonétique de la phrase "Une rose fanée". Le double branchement a été indiqué en pointillé.*

Résultats pour le dictionnaire SLC

Le **dictionnaire SLC**, développé pour le traitement de la langue naturelle écrite au CNET, contient à peu près 50000 entrées complétées par leur formes fléchies.

La transcription proposée par le dictionnaire est une transcription phonémique avec des variantes de prononciation (prononciation plus soutenue contenant des [ə] élidables et prononciation plus coarticulée sans [ə] élidable; voyelles à deux apertures ...). L'inconvénient majeur de la phonétisation fournie par ce dictionnaire est la non indication des phonèmes finaux latents qui peuvent apparaître en cas de liaison (ex: **"petit homme" [p.ə.t.i.ɔ.m]** au lieu de **[p.(ə).t.i.t.ɔ.m], "des enfants" [d.e.ã.f.ã]** au lieu de **[d.e.z.ã.f.ã]**...). Quand cela est possible, le dictionnaire propose des variantes de prononciation, exemple **"semaine"** a deux phonétisations: **[s.ə.m.ɛ.n]** et **[s.m.ɛ.n]**. Malheureusement, ce traitement n'est pas systématique. Ainsi, certains mots ont reçu une seule phonétisation "longue", sans élision du schwa (par exemple **"petit"** est phonétisé **[p.ə.t.i]** et pas **[p.t.i]**). D'autres en ont également reçu une seule mais "courte", contenant une élision définitive du schwa (exemple : **"rapidement"** devient **[r.a.p.i.d.m.ã]** et pas **[r.a.p.i.d.ə.m.ã]**) **"passeport"** donne **[p.a.s.p.ɔ.r]** et pas **[p.a.s.ə.p.ɔ.r]** ou **"médecin"** est phonétisé **[m.e.d.s.ɛ̃]** et pas **[m.e.d.ə.s.ɛ̃]**). En cas de phonétisation "courte", il nous est difficile, voire impossible par un traitement phonologique simple, appliqué après la phonétisation du mot, de restituer le son élidé. Cela peut causer des difficultés au niveau de la segmentation du signal puisque beaucoup de locuteurs, surtout dans une élocution relativement lente, de surcroît en mots isolés, vont "sur articuler" le mot, gardant dans leur prononciation le schwa supplémentaire.

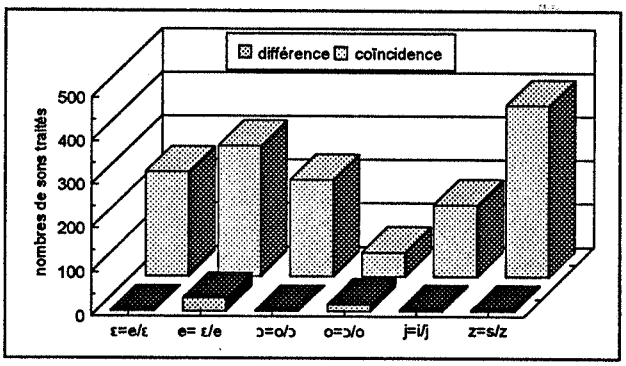


Figure 3: Comparaison entre phonétisation automatique (dictionnaire SLC) et manuelle.

Le nombre de substitutions de phonèmes était très faible: 78 sur 5461 phonèmes traités. Comme cela apparaît sur la figure 3, les substitutions les plus courantes se sont produites entre les voyelles d'aperture variable comme [ɛ] versus [e] ou [ɔ] versus [o]. Ainsi, par exemple, pour la voyelle [e] le consensus entre la transcription manuelle et automatique s'est élevé à 365 cas et dans 32 cas le [e] de la phonétisation manuelle correspondait à [ɛ] de la phonétisation automatique. Quelques substitutions sont intervenues également au niveau du changement de voisement des consonnes.

Résultats pour le dictionnaire BDLEX

Le dictionnaire du GDR-PRC "Communication Homme-Machine", **BDLEX,** est plus proche par sa conception des besoins de la reconnaissance automatique de la parole que ne l'est le dictionnaire SLC. En effet, la transcription phonétique proposée par ce dictionnaire contient des formes phonétiques longues (sans élision des schwas) ainsi que les phonèmes latents (phonèmes de liaison). Le nombre d'entrées est à peu près de 20000 mots complétés par leurs formes fléchies. Le dictionnaire BDLEX utilise une transcription phonémique.

D'une façon générale, l'inconvénient majeur de ce dictionnaire est sa petite taille. Un des reproches que l'on peut faire à ce dictionnaire est l'absence de l'indication des latences des schwas à l'intérieur des mots. Si l'on considère systématiquement ces schwas comme élidables, on peut introduire des variantes de prononciation hautement improbables en français (par exemple: **[r.ɛ.g.l.m.an]** au lieu de **[r.e.g.l.ə.m.an]**, **[p.r.m.j.ɛ.r]** au lieu de **[p.r.ə.m.j.ɛ.r]**, **[r.p.r.i.z]** au lieu de **[r.ə.p.r.i.z]**; **[s.a.t.i.s.f.s.ã]** au lieu de **[s.a.t.i.s.f.ə.s.ã]**; **[r.s.t.r.y.k.t.y.r.e]** au lieu de **[r.ə.s.t.r.y.k.t.y.r.e]**...) et augmenter la complexité des décisions lors de la reconnaissance.

Au niveau phonémique, le dictionnaire BDLEX ne maintient pas la distinction entre **[ɛ]** et **[e]**. Le traitement des autres voyelles à double aperture (ouverte et fermée) se fait également d'une façon sommaire: en syllabe ouverte la voyelle devient fermée et en syllabe fermée la voyelle devient ouverte. Ainsi, par exemple, on trouve **[k.o.m.ɛ.r.s]** au lieu de **[k.ɔ.m.ɛ.r.s]**; **[k.o.m.ã.d]** au lieu de **[k.ɔ.m.ã.d]**; **[r.e.ʒ.j.o.n.a.l]** au lieu de **[r.e.ʒ.j.ɔ.n.a.l]** ou **[o.r.ɛ.r]** au lieu de **[ɔ.r.ɛ.r]**, **[a.m.u.r.œ.z]** au lieu de **[a.m.u.r.ø.z]**... Une autre difficulté d'utilisation de ce dictionnaire pour la reconnaissance automatique de la parole, réside dans son traitement des voyelles nasales en position finale des mots. Dans ces contextes, les voyelles nasales sont réécrites sous leur forme de liaison, c'est-à-dire en voyelles orales et en consonnes latentes nasales. Ainsi la phonétisation du mot **"divin"** est **[d.i.v.i.(n)][1]**, le mot **"prochain"** est phonétisé comme **[p.r.o.ch.ɛ.(n)]** ou le mot **"lion"** comme **[l.j.ɔ.(n)]**...

Le dictionnaire BDLEX prévoit un traitement de liaisons sur les frontières des mots. Mais le plus souvent c'est le graphème final qui est proposé comme phonème latent. Ainsi, par exemple, le mot **"bal"** sera transcrit sans schwa latent **[b.a.l]** tandis que le mot

[1] (n) indique la consonne latente de liaison.

"**balle**" avec un schwa latent **[b.a.l.(ə)]** bien que dans les deux cas la réalisation d'un schwa final soit équiprobable au niveau acoustique. Au même titre, le mot "**vue**" sera phonétisé **[v.y.(ə)]** et l'adjectif cardinal "**sept**" sera phonétisé **[s.ɛ.t]**, alors qu'au niveau acoustique on a une probabilité considérable de trouver l'inverse, c'est-à-dire **[s.ɛ.t.ə]** et **[v.y]**.

Le nombre d'insertions de phonème, réalisé par le dictionnaire BDLEX est de 406[2]. Le nombre d'omissions s'élève à 76. Le nombre de substitutions est également élevé: 382 substitutions pour 5549 phonèmes traités. La figure 4 retrace les différences et les coïncidences entre la phonétisation manuelle et la phonétisation fournie par le dictionnaire BDLEX.

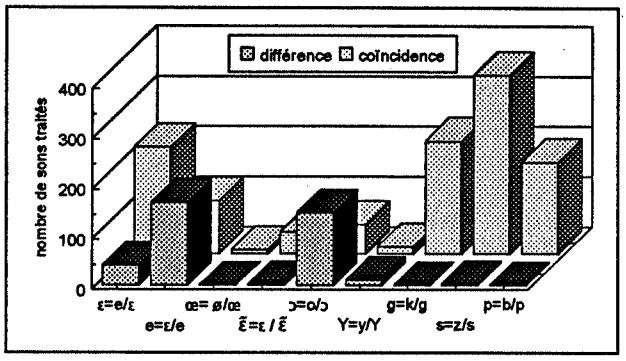


Figure 4 : Comparaison entre phonétisation automatique (dictionnaire BDLEX) et manuelle

Résultats pour le phonétiseur par règles

Le système de **phonétisation par règles** a été élaboré initialement pour la synthèse de la parole. La phonétisation se fait au niveau phonémique et une modification a été apportée pour faire la distinction entre les **[ə]** instables à l'intérieur des mots et le schwa latent de la fin de mot après une consonne.

Le nombre d'omissions s'élève pour le phonétiseur par règles à 268. La cause majeure de ce grand nombre d'omissions est le non-traitement des phonèmes de liaison. En effet, étant donnée sa destination originale (synthèse de la parole) le système de phonétisation par règles est capable de prévoir les phonèmes latents uniquement en cas de connaissance du contexte immédiat. Comme en reconnaissance de la parole la phonétisation des mots a lieu avant la compilation des informations syntaxiques, la consonne de liaison ne peut pas être déterminée par le phonétiseur.

Le nombre d'insertions de phonèmes était de 430 pour le phonétiseur par règles. Il s'agissait essentiellement (dans 408 cas) d'insertions de l'attribut de latence du **[ə]** instable. Comme cela apparaît de la figure 5, le nombre de substitutions était peu élevé: dans 81 cas pour 6823 phonèmes traités, la phonétisation automatique s'écartait de la phonétisation manuelle.

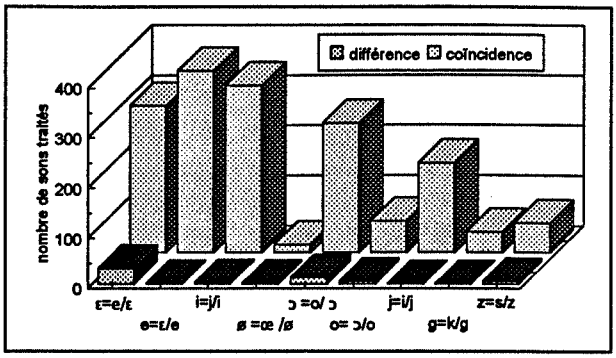


Figure 5: Comparaison entre phonétisation automatique (phonétiseur par règles) et manuelle.

Comparaison des trois phonétiseurs

Pour la reconnaissance automatique de la parole, il est raisonnable de considérer comme plus grave de substituer des phonèmes ou d'en omettre que d'insérer des phonèmes latents. En effet, en substituant un son par un autre, on s'éloigne de la transcription considérée comme référence. L'omission d'un son latent ou simplement de l'attribut de latence d'un son peut introduire des problèmes de modélisation au niveau des HMM. Par contre, l'insertion d'un son latent ou le caractère non stable d'un son ne fera qu'introduire un double branchement au niveau des HMM en proposant des variantes de prononciation. Pour cette raison, nous avons basé notre comparaison théorique sur les mesures de substitutions et d'omissions.

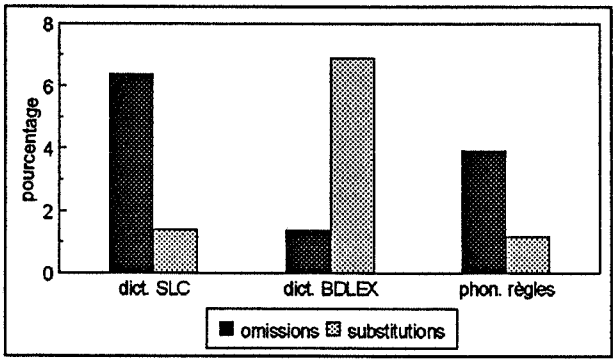


Figure 6: Pourcentage d'omissions et de substitutions par rapport au nombre total des phonèmes traités.

Comme cela apparaît sur la Figure 6, c'est le phonétiseur par règles qui termine en moyenne le mieux cette comparaison théorique. Il se place le mieux quant au pourcentage de substitutions mais obtient également d'assez bons résultats pour le pourcentage d'omissions.

[2] Dans 352 cas il s'agissait d'insertions de **[e]** latents en fin de mot ou de phrase.

PERFORMANCES DE RECONNAISSANCE

Les tests de reconnaissance ont été effectués avec le système de reconnaissance Phil90 [9] basé sur la modélisation markovienne. Les unités modélisées par les HMM étaient des allophones (modèles de phonèmes en contexte). Les corpus traités étaient les suivants: **chiffres, nombres, Trégor, 500_mots et trains.**

L'initialisation des modèles des différentes bases a été réalisée à partir d'un modèle appris sur la base des allophones transcrite manuellement. Comme les dictionnaires SLC et BDLEX ne phonétisaient pas tous les mots des corpus, les tests de reconnaissance ont été réalisés sur l'intersection des corpus phonétisés (4 premières lignes de la table I ci-dessus). Les bases de données ont été divisées en deux parties, la première servait pour apprendre les paramètres du modèle et la deuxième pour mesurer la performance des modèles appris.

Comme le phonétiseur par règles phonétisait 100% des corpus, nous avons également comparé sa performance avec celle obtenue par la phonétisation manuelle sur l'intégralité des corpus (les 2 dernières lignes de la table I ci-dessus: man/tot et règles/tot).

Chaque phonétiseur a été appliqué sur chaque corpus. Des tests de reconnaissance ont été effectués pour comparer les performances de Phil90 utilisant la description des bases de données fournie par les 3 phonétiseurs et par la transcription manuelle. De ces résultats 1 était significativement meilleur et 8 (2 sur l'intégralité des corpus et 6 sur l'intersection des corpus) significativement plus mauvais (se trouvaient à l'extérieur de l'intervalle de confiance) que les scores obtenus par la transcription manuelle. La table I présente un extrait des résultats obtenus. Elle contient le taux d'erreur obtenu pour la phonétisation manuelle avec son intervalle de confiance ainsi que tous les taux d'erreurs des phonétisations automatiques qui se trouvent à l'extérieur de cet intervalle (indiqués en gras).

Table I : *Intervalle de confiance et extrait de taux d'erreurs de reconnaissance en %.*

	App. indépendant de l'application				Dépend.[3]
	trains	Trégor	nombres	500 mots	500 mots
manuelle	36,89 ±2,31	2,62 ±0,31	8,75 ±0,65	16,04 ±1,06	6,88 ±0,73
SLC	*36,59*	*2,96*	*9,99*	*16,54*	*7,11*
BDLEX	*34,61*	*2,62*	*9,78*	*17,36*	*7,70*
phon_règ.	*33,77*	*2,61*	*10,11*	*16,01*	*7,25*
man/tot	31,29 ±1,06	2,49 ±0,27	8,75 ±0,65	15,58 ±1,02	6,94 ±0,71
règles/tot	*33,11*	*2,77*	*10,30*	*16,57*	*6,64*

Sur les corpus de comparaison des 3 phonétiseurs, c'est le phonétiseur par règles qui fournit le meilleur résultat. Le score obtenu par cette description (sur le corpus commun) s'est trouvé une fois significativement plus mauvais et une fois significativement meilleur que celui obtenu sur les mêmes tests par la transcription manuelle.

Comme cela apparaît sur la Table I, les écarts significativement plus mauvais se trouvent essentiellement dans la partie des tests où l'apprentissage était indépendant de l'application. Dans la partie des tests avec apprentissage dépendant de l'application, les paramètres s'ajustaient lors de l'apprentissage aux différentes phonétisations, et l'écarts s'effaçait parmi elles. En reconnaissance automatique de la parole, un de nos objectifs est de rendre l'apprentissage indépendant de l'application. Par conséquent les résultats obtenus sur cette partie des tests s'avèrent pour nous très importants.

CHOIX DU PHONETISEUR ET ADAPTATION POUR LA RECONNAISSANCE

A la suite des tests réalisés dans cette étude, notre choix s'est porté sur l'utilisation du phonétiseur par règles pour la reconnaissance automatique de la parole. Ce choix a été motivé par les points suivants:

- ***par sa "ressemblance" avec la phonétisation manuelle***
- ***par les bonnes performances de reconnaissance obtenues avec ce phonétiseur***
- ***par sa couverture de phonétisation***
- ***par la richesse des informations disponibles dans ce système.***

Bien entendu, il est nécessaire d'adapter ce système de phonétisation pour les besoins de la reconnaissance automatique de la parole. Comme cela a été évoqué plus haut, la phonétisation du vocabulaire pour la reconnaissance de la parole est réalisée sous forme de mots isolés. Pour cette raison, il nous est impossible d'utiliser l'analyseur grammatical (disponible dans le phonétiseur par règles), dont le rôle est de lever l'ambiguïté des sens parmi plusieurs mots en cas de conflit. Par conséquent, pour la reconnaissance de la parole, le phonétiseur doit travailler en mode interactif. Cela veut dire qu'en cas d'homographes hétérophones c'est à l'utilisateur de lever l'ambiguïté au moment de la phonétisation.

Une des richesses de ce phonétiseur est son dictionnaire des locutions [10]. En effet, en synthèse de la parole (pour laquelle ce système de phonétisation a été initialement conçu) il est important de connaître les locutions de sorte qu'au niveau prosodique, l'on n'introduise jamais de pause entre leurs éléments. En reconnaissance de la parole il est également important de reconnaître certaines locutions, car par la forte cohésion qui existe entre leurs éléments, le locuteur fera

[3] Dépend. = Apprentissage dépendant de l'application.

sonner des phonèmes latents qui normalement ne se prononcent jamais. Ainsi par exemple, les consonnes finales (muettes) des mots tels que **"nuit", "pied", "mot", "pont"...** ne seront jamais prononcées, sauf dans la locution **"nuit-et-jour" [n.Y.i.t.e.ʒ.u.r], "pied-à-terre" [p.j.e.t.a.t.ɛ.r], "mot-à-mot" [m.ɔ .t.a.m.o], "temps-en-temps" [t.ã.z.ã.t.ã], "pont-aux-ânes" [p.ɔ̃.t.o.z.a.n]...**

Un autre avantage du phonétiseur par règles est sa distinction entre les [ə] non-stables et les [ə] permanents à l'intérieur du mot. Le système utilise une analyse du contexte immédiat des [ə] instables, afin de les réécrire en phonème non-élidable au voisinage d'un cluster consonantique[4]. On pourrait ainsi éviter la modélisation de mots imprononçables en français (comme **[o.t.r.m.ã], [r.s.t.r.y.k.t.y.r.e], [r.d.r.ɛ.s.e], [f.s.a.b.l]...**) et réduire la complexité des choix lors de la reconnaissance de la parole.

Dans cette étude, nous n'avons pas abordé au niveau phonémique le problème d'assimilation de différents traits d'articulation par les phonèmes avoisinants. La raison de cela est la suivante: il est plus réaliste de traiter l'assimilation des traits phonétiques (ex: changements de trait de voisement, exportation de trait nasal...) au niveau acoustique, dans la modélisation même des allophones. Ainsi, par exemple, en cas d'élision du schwa dans un mot tel que **"médecin" [m.e.d.ə.s.ɛ̃]** qui devient **[m.e.t.s.ɛ̃]**, nous n'avons pas besoin de règle de réécriture de **[d]** en **[t]** dans un contexte sourd, car les "boîtes" acoustiques des allophones vont autoriser un passage par une cible complètement ou partiellement dévoisée quand le contexte droit immédiat d'une consonne voisée se trouve être une consonne sourde [11]. Grâce à des possibilités d'inter-branchements entre les états du modèle de Markov, modélisant les cibles des sons, nous nous plaçons plus près de la réalité du signal, puisque nous admettons que des phénomènes tels que l'assimilation des traits puissent être partiels ou globaux alors même qu'en phonologie ces changements sont considérés le plus souvent comme des changements tout-ou-rien.

CONCLUSION

Cette étude sur la comparaison de trois phonétiseurs nous a permis de choisir le phonétiseur le mieux adapté et d'affiner nos exigences en ce qui concerne le travail d'un phonétiseur dans le domaine de la reconnaissance automatique de la parole.

Les phonétisations testées dans cette étude ont été réalisées par trois phonétiseurs au niveau phonémique (transcription large: à peu près 16 voyelles, 3 semi-voyelles et 18 consonnes ont été utilisées pour décrire la prononciation des mots). Le système de reconnaissance utilisé transformait cette description en modélisation par allophones. Une partie des traitements que l'on peut considérer phonémique comme par exemple l'effacement de consonnes de liaison en contexte consonantique) ou la propagation des traits assimilés par le son à cause de son entourage (assimilation partielle ou globale) a été prise en charge par des règles introduites soit dans la bibliothèque des HMM soit directement dans la description acoustiques des modèles des allophones.

Le phonétiseur, retenu pour la reconnaissance automatique de la parole, est le **phonétiseur par règle**. Ce choix a été justifié par ses performances quant à la couverture des corpus phonétisés (100%), la "ressemblance" avec la phonétisation manuelle et les bonnes performances de reconnaissance obtenue avec ce phonétiseur. De plus, le phonétiseur par règles, adapté pour la reconnaissance de la parole, fournit des formes longues des mots ([ə] instable non-élidé et phonèmes de liaison), vérifie la "stabilité" des [ə] en fonction de leur contexte immédiat et utilise son lexique de locutions pour prédire certaines liaisons consonantiques, autrement non réalisées.

[4] Cette fonctionalité n'a pas été utilisée lors de nos tests.

Références

[1] Marty F.: *Trois systèmes informatiques de transcription phonétique et graphémique, **"Le Français moderne"***, Décembre, 1992.

[2] Ombrouck J. M.: ***"Extraction automatique d'informations lexicales à partir d'un dictionnaire de langue"***, NT/LAA/SLC/304, 1988;

[3] ***"Base de données lexicales du français écrit et parlé (BDLEX) du GDR-PRC "Communication Homme-Machine" - France.***

[4] Bartkova K. : ***"Choix des corpus pour la reconnaissance de la parole pour la modélisation des allophones";*** DT 1048/LAA/TSS/RCP, Novembre 1990.

[5] L.Warnant : ***"Dictionnaire de la prononciation française dans sa norme actuelle"***, Duculot, Paris, 1987.

[6] Prunet J-F : ***"L'effacement en français populaire parisien";*** Mémoire de maîtrise, Université Paris 7, 1981.

[7] Lacheret-Dujour A.: ***"Contribution à l'analyse de la variabilité phonologique pour le traitement automatique de la parole continue multilocuteur"***, Thèse de Doctorat, Université Paris 7, 1990.

[8] Fletcher J. : *Some effects of tempo change on Timing in French: considerations for speech synthesis-by-rule systems,* **"European Conference on Speech Technology"**, Edinburgh, 2-4 Septembre 1987.

[9] Jouvet D., Bartkova K. & Monné J. : *On the modelization of allophones in an HMM based speech recognition system;* **Eurospeech'91**, Genova, Italy, 24-26 Septembre 1991.

[10] Larreur D., Emerard E., Marty F. : *"Linguistic and Prosodic Processing for a Text-to-Speech Synthesis System"*, **EUROSPEECH**, vol 1, pp. 510-513, Paris, 1989.

[11] Bartkova K. & Jouvet D.: *Modelization of allophones in a speech recognition system;* **ICPhS**, Aix-en-Provence, France, 19-24 Août, 1991.

UN ALGORITHME POUR L'ESTIMATION DES INSTANTS DE FERMETURE DE LA GLOTTE À L'AIDE DE LA DÉTECTION SÉQUENTIELLE DES RUPTURES DU SIGNAL DE PAROLE

C. MURGIA & G. FENG

Institut de la Communication Parlée
INPG/ENSERG, Université Stendhal, URA CNRS N° 368,
B.P. 25, 38040 GRENOBLE CEDEX 9 FRANCE

Résumé

Une estimation précise de la fréquence fondamentale est un aspect essentiel d'un grand nombre de systèmes de traitement de la parole. Au fil des années, un certain nombre d'algorithmes capables de détecter les instants de fermeture de la glotte ont été proposés. Deux nouveaux algorithmes basés sur la détection de ruptures dans les propriétés statistiques des signaux sont particulièrement intéressants, toutefois plusieurs problèmes existent pour les nasales, les fricatives et plosives voisées (Moulines et Di Francesco, 1990). Nous présentons ici un nouveau et efficace algorithme de détection des instants de fermeture de la glotte pour la mesure de la micromélodie du signal de parole. Ce système basé sur la détection de ruptures est capable de traiter tous les sons voisés.

Mots-clés: fermeture de la glotte, micromélodie, détection automatique de ruptures.

I. INTRODUCTION

La fréquence fondamentale est un paramètre important du signal de parole. Son estimation précise constitue un des éléments les plus importants dans un grand nombre de systèmes de codage et de synthèse. Elle peut également être utile et parfois essentielle pour l'identification des locuteurs, car ses variations véhiculent des informations phonétiques et linguistiques concernant le locuteur.

La mesure de la fréquence fondamentale est une tâche difficile pour plusieurs raisons: non-stationnarité du signal, variabilité de la parole, irrégularité de la source de voisement. De nombreux algorithmes ont été proposés afin de résoudre ce problème, cependant la qualité des résultats obtenus n'est pas toujours satisfaisante. En effet, une grande partie de ces techniques fournissent une estimation de la valeur moyenne du fondamental de la trame analysée (Sondhi, 1968; Noll, 1970; Markel, 1972; Ross et al., 1974; Wise et al., 1976; Martin, 1981). Au fil des années, un certain nombre de méthodes capables de détecter les instants de fermeture de la glotte ont été mises au point. Ces techniques, conçues essentiellement pour apporter une solution au problème du filtrage inverse, présentent toutes des difficultés d'application spécifiques dues soit à l'importante complexité des calculs, soit à des hypothèses trop strictes faites sur le signal à analyser (Young, 1965; Strube, 1974; Ananthapadmanabha et Yegnanarayana, 1979; Cheng et O'Shaughnessy, 1989).

Récemment deux algorithmes (maximisation séquentielle d'un rapport de vraisemblance et détection des ruptures de convexité dans la fonction de vraisemblance) basés sur les changements des propriétés statistiques du signal ont été développés, mais les résultats obtenus pour les nasales, les fricatives voisées et les plosives voisées ne sont pas totalement satisfaisants (Moulines et Di Francesco, 1990).

Dans cette communication, nous proposons un nouvel algorithme pour la détection des instants de fermeture de la glotte à partir de la forme d'onde acoustique. Cet algorithme est basé sur la détection des changements rapides dans le signal de parole engendrés par la fermeture des cordes vocales. Nous utilisons la modélisation adaptative pour améliorer les performances de la statistique sur les sons voisés. Nous avons mis au point un algorithme capable d'extraire l'information de micromélodie pour tous les sons voisés.

II. RAPPEL DU PRINCIPE DE LA DÉTECTION DE RUPTURES

Le but de la détection automatique de ruptures est de repérer les instants qui correspondent aux changements spectraux et énergétiques importants du signal. La technique que nous utilisons est basée sur une mesure de distance entre deux modèles. En effet, le signal est observé sur deux échelles de temps : une fenêtre d'observation à long terme et une autre à court terme. Ces deux tranches de signal sont modélisées à l'aide de deux modèles LPC et, pour chaque modèle, on calcule la valeur de la densité de probabilité (ddp) en utilisant les résidus LPC. Un test statistique est ensuite utilisé pour mesurer la distance entre ces deux ddp. Les observations avant rupture suivent le modèle long terme (hypothèse H_0) et, après rupture, le modèle court terme (hypothèse H_1). Le test utilisé est la somme cumulée du rapport de vraisemblance a posteriori des deux distributions de probabilité, corrigé par la divergence de Kullback (Basseville et Benveniste, 1983). Le cas simple de la modélisation autorégressive et des résidus gaussiens conduit à l'expression suivante de l'incrément du test :

$$w_n = \frac{1}{2}\left[\frac{2\, e_n^0\, e_n^1}{\sigma_1^2} - \left[1 + \frac{\sigma_0^2}{\sigma_1^2}\right]\frac{(e_n^0)^2}{\sigma_0^2} + \left[1 - \frac{\sigma_0^2}{\sigma_1^2}\right]\right]$$

avec e_n^i (i = 0,1) les résidus des modèles long et court terme, et σ_i^2 les variances des résidus des modèles long et court terme. On peut démontrer que la somme cumulée de ces incréments oscille autour de zéro dans les zones stationnaires et chute lorsque se produit un changement des caractéristiques du signal.

III. DÉTECTION DES INSTANTS DE LA FERMETURE DE LA GLOTTE : APPORT DE LA MODÉLISATION ADAPTATIVE

Le test de divergence a été initialement appliqué au signal de parole pour la segmentation automatique dans le cadre de la reconnaissance de la parole (André-Obrecht, 1988). Néanmoins, on peut distinguer dans le signal de parole deux types de ruptures : les macroruptures, qui correspondent aux transitions entre deux zones du signal relativement stationnaires et les microruptures liées à la fermeture brusque de la glotte, phénomène qui se produit à chaque période fondamentale dans les sons voisés. La somme cumulée réagit à chaque microrupture par un saut dû au pic négatif de son incrément. La chute de la statistique se produit en correspondance avec la rupture. En effet, lorsque le signal se trouve sur une zone stationnaire (comprise entre deux instants de fermeture de la glotte) les modèles court terme et long terme sont proches; à chaque fois qu'une rupture se produit, elle sera prise en compte dans la mémoire des deux modèles. Toutefois, à cause de la différence de taille des deux fenêtres d'analyse, elle aura plus d'importance sur le modèle court terme que sur le long terme : les deux modèles divergent et cette différence est indiquée par la statistique avec un saut négatif.

Cette méthode peut fournir donc un outil très performant pour la détection de l'instant de fermeture de la glotte. Cependant, le test statistique ne réagit pas aux microchangements spectraux pour certains types de sons voisés (Moulines et Di Francesco, 1990).

La modélisation adaptative a été initialement utilisée pour réduire le comportement dissymétrique de la statistique (Feng et al., 1991), toutefois nous avons pu constater que son introduction améliore la sensibilité du test aux microruptures pour tous les sons voisés. Nous avons donc utilisé la modélisation de Burg avec coefficient d'oubli pour estimer le modèle long terme et l'algorithme *prewindow* pour le modèle court terme (Friedlander, 1982). Comme notre objectif est la détection des microruptures et non des macroruptures nous avons effectué une étude systématique pour déterminer les paramètres des modèles adaptatifs (coefficients d'oubli et ordres des filtres LPC). Les paramètres retenus nous ont consenti d'obtenir des pics de l'incrément très nets et profonds, ce qui permet une détection précise des instants de fermeture de la glotte et une estimation précise des périodes fondamentales (Murgia, 1992). Le rôle des filtres LPC est ici de réduire la redondance du signal. Il n'est donc pas nécessaire d'utiliser un ordre trop élevé. Nous avons donc choisi pour les deux filtres un ordre égal à 4 ($M_1 = M_0 = 4$). En ce qui concerne les coefficients d'oubli, nous les avons choisis de façon à avoir des fenêtres exponentielles plus petites que à celles utilisées dans la détections des macrochangements (Feng et al., 1991) : $w_0 = 0.953$ pour le modèle long terme et $w_1 = 0.939$ pour le modèle court terme. Il est souhaitable d'avoir des fenêtres de longueur inférieure au *pitch*. Dans ce cas la prise en compte dans les deux filtres de la rupture sera un événement (*epoch*) plus important et les pics de l'incrément de la statistique seront plus nets et profonds.

Pour illustrer la performance de la méthode et la validité des paramètres sélectionnés, nous montrons le comportement de la statistique calculée à partir des modèles adaptatifs pour les sons [a], [i], [m], [z] et [da]. Les signaux sont échantillonnés à 16 kHz et codés sur 16 bits. Un cas relativement facile à interpréter est celui du son [a] (figure 1). Sur le signal temporel les oscillations amorties de la réponse du conduit vocal aux impulsions glottiques sont parfaitement lisibles et l'on constate une bonne correspondance entre le début de la période et le pic de l'incrément de la statistique.

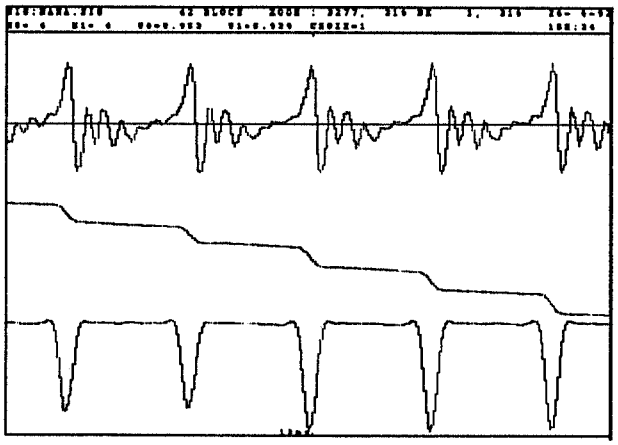


Fig. 1. Allure de la statistique et de son incrément sur le son [a]. Du haut en bas : le signal, le test "somme cumulée", l'incrément de la statistique.

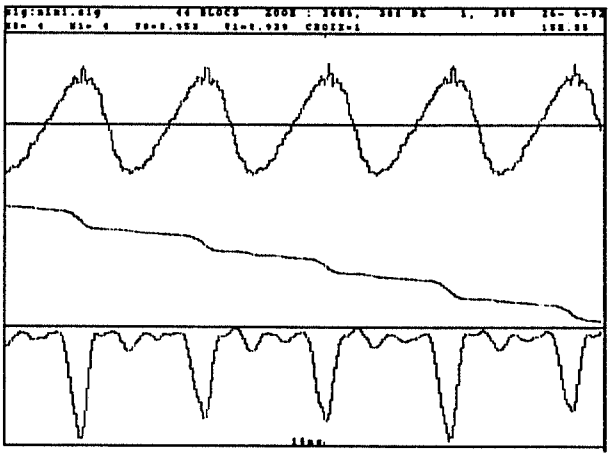


Fig. 2. Allure de la statistique et de son incrément sur le son [i]. Même légende que pour la figure 1.

La figure 2 illustre le comportement de la statistique sur le son [i]. Presque tous les algorithmes basés sur les résidus LPC ont des difficultés pour ce type de signal. En effet, dans le cas où le premier formant est très proche du fondamental, le résidu LPC ne présente pas une périodicité nette (Hess, 1983). Par ailleurs, il est très difficile de localiser l'instant d'excitation à partir de l'observation directe du signal, car celui-ci est presque sinusoïdal. En revanche, la statistique fournit une série de pics périodiques qui nous semblent indiquer correctement les instants de fermeture de la glotte.

La figure 3 montre le cas de [m] : le test réagit tout à fait correctement pour les nasales.

Examinons maintenant le cas de [z], pour lequel l'extraction du fondamental est particulièrement difficile. En effet, pour les fricatives voisées il existe deux types d'excitations en même temps : la vibration des cordes vocales et un bruit de friction dû à l'écoulement de l'air au travers du rétrécissement du conduit vocal. De plus, pour ces sons, les cordes vocales ne sont jamais complètement fermées (Moulines et Di Francesco, 1990). En figure 4, on peut constater que la statistique fournit des pics périodiques, ce qui montre la validité de la méthode pour les fricatives voisées.

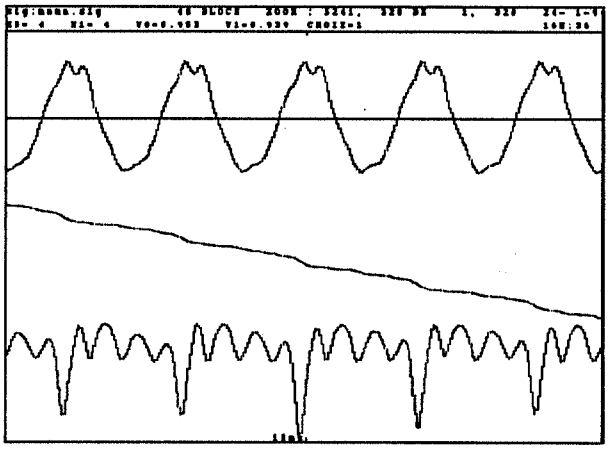


Fig. 3. Allure de la statistique et de son incrément sur le son [m]. Même légende que pour la figure 1.

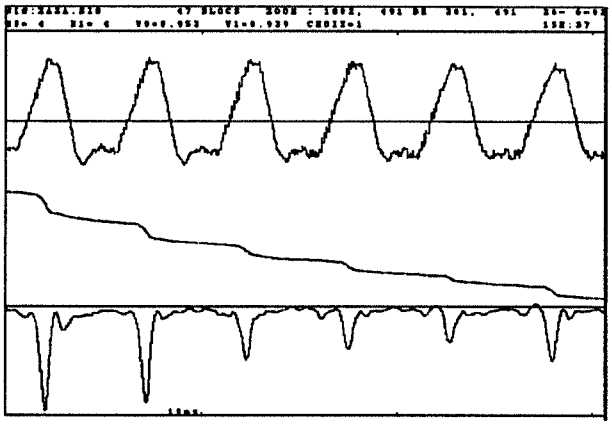


Fig. 4. Allure de la statistique et de son incrément sur le son [z]. Même légende que pour la figure 1.

La figure 5 montre le comportement de la statistique pour la transition [da]. Pour ces sons, la difficulté majeure se trouve dans la zone où le bruit d'occlusion se superpose à la source de voisement. Dans ce cas, les variations spectrales introduites par ce bruit peuvent masquer les ruptures liées à la fermeture de la glotte. Cela peut constituer une source d'erreurs. Une méthode basée sur le filtrage du signal est en cours de développement.

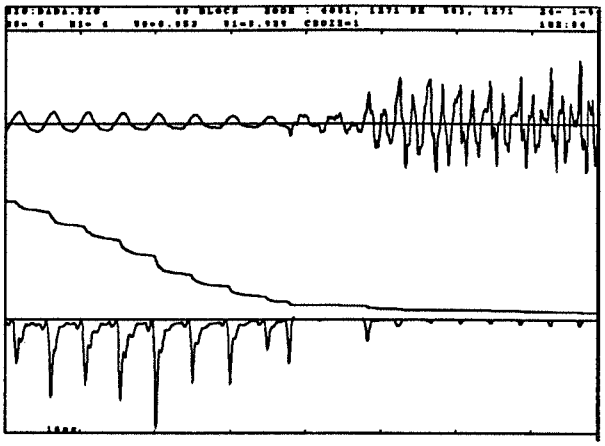


Fig. 5. Allure de la statistique et de son incrément sur le son [da]. Même légende que pour la figure 1.

IV. UN NOUVEL ALGORITHME DE DÉTECTION

Nous avons mis au point un algorithme de mesure de la micromélodie du signal de parole basé sur la détection des instants de fermeture de la glotte localisés par les pics de l'incrément de la statistique. Un détecteur de voisement sélectionne les segments voisés pour la mesure de la fréquence fondamentale. L'algorithme utilisé est basé sur la mesure de l'énergie, du taux de passage par zéro et du rapport des deux premiers coefficients d'autocorrélation calculés toutes les 10 ms.

Une fois que les zones voisées sont localisées, il reste à résoudre le problème de la sélection du pic de l'incrément de la statistique correspondant aux instants de fermeture de la glotte. En effet, à l'intérieur d'une période, l'incrément de la statistique présente toujours plusieurs pics à cause des diverses variations spectrales et temporelles. On considère que le pic négatif le plus profond correspond à la fermeture de la glotte, parce que cet événement engendre sur le signal les variations spectrales plus significatives. Toutefois la notion de "plus profond" est une notion locale. Donc, afin de détecter le pic nous découpons le signal en fenêtres de recherche de taille variable. La taille de la fenêtre est calculée en fonction de la moyenne pondérée des mesures des périodes précédentes ($<T_0>$). En effet, on sait que la variation du fondamental peut atteindre ±40% à l'intérieur d'un segment voisé (Hess, 1983). Le pic indiquant l'instant de fermeture de la glotte doit donc se trouver dans une zone définie par la position plus probable du pic recherché ±40% de $<T_0>$, où la position

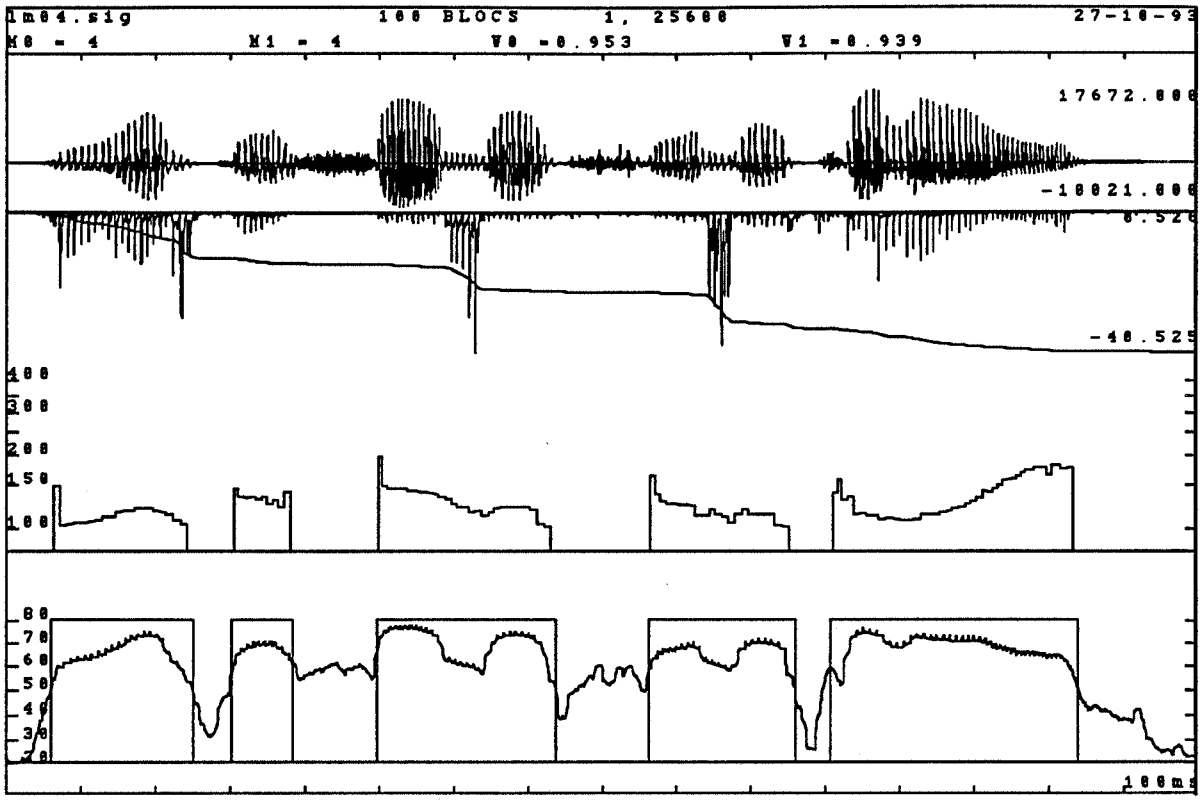


Fig. 6. Résultats de mesure de F_0. De haut en bas : le signal de parole pour la phrase *Vous poussez des cris de colère?*, l'incrément du test statistique, le test somme cumulée, le contour de F_0 en Hz, la décision voisé/non-voisé et l'énergie du signal en dB.

plus probable est donnée par la position du pic précédemment sélectionné plus <T_0>. Des marqueurs indiqueront la position des pics associés aux instants de fermeture de la glotte. L'utilisation de la moyenne des estimations précédentes nous permet de limiter les risques de propagation des erreurs de mesure et la pondération nous permet de donner plus d'importance aux estimations plus récentes qui sont les plus proches de la valeur à mesurer. Enfin, on mesure la fréquence fondamentale comme l'inverse de la distance entre deux marqueurs successifs.

V. ÉVALUATION DES RÉSULTATS

Une première évaluation de notre algorithme a été effectuée sur un corpus de 24 phrases françaises, prononcées par six locuteurs (trois femmes et trois hommes, signaux échantillonnés à 16 kHz et codés sur 16 bits). Les résultats obtenus montrent un bon fonctionnement de l'algorithme : le fondamental de la plupart des segments voisés est correctement mesuré. Les nasales, les fricatives voisées et les plosives voisées ne semblent pas poser des problèmes particuliers.

La figure 6 montre les résultats pour la phrase *Vous poussez des cris de colère?*. On peut notamment remarquer l'efficacité du système pour la fricative voisée [v] et pour la plosive voisée [d]. Pour évaluer la qualité de notre algorithme nous comparons les résultats avec ceux d'un algorithme classique basé sur la détection des passages par zéro. Dans ce dernier le signal de parole est d'abord filtré par un filtre passe-bas dont la fréquence de coupure est préalablement choisie en fonction du locuteur afin de ne maintenir que le fondamental (Boë et Rakotofiringa, 1971, Antoniades, 1993). Nous utilisons cette méthode comme référence car elle fournit une bonne estimation du fondamental presque période par période. De plus, l'effet du filtrage assure un bon lissage du résultat. On peut constater que les mesures de notre algorithme sont globalement cohérentes avec la référence (figure 7). Une agitation plus grande se trouve dans notre résultat. Cela est dû à la structure intrinsèque de l'algorithme. En effet, ayant pour but de détecter les instants de la fermeture de la glotte, celui-ci doit avoir une sensibilité relativement grande. Il faut indiquer que la courbe de F_0 n'a subi aucun post-traitement ou lissage. Toutefois, dans quelques zones de transition, comme celles entre certaines occlusives voisées et les voyelles, des imprécisions de mesure peuvent se produire. En effet, dans ces cas le test statistique réagit plutôt aux macrochangements parce que ceux-ci

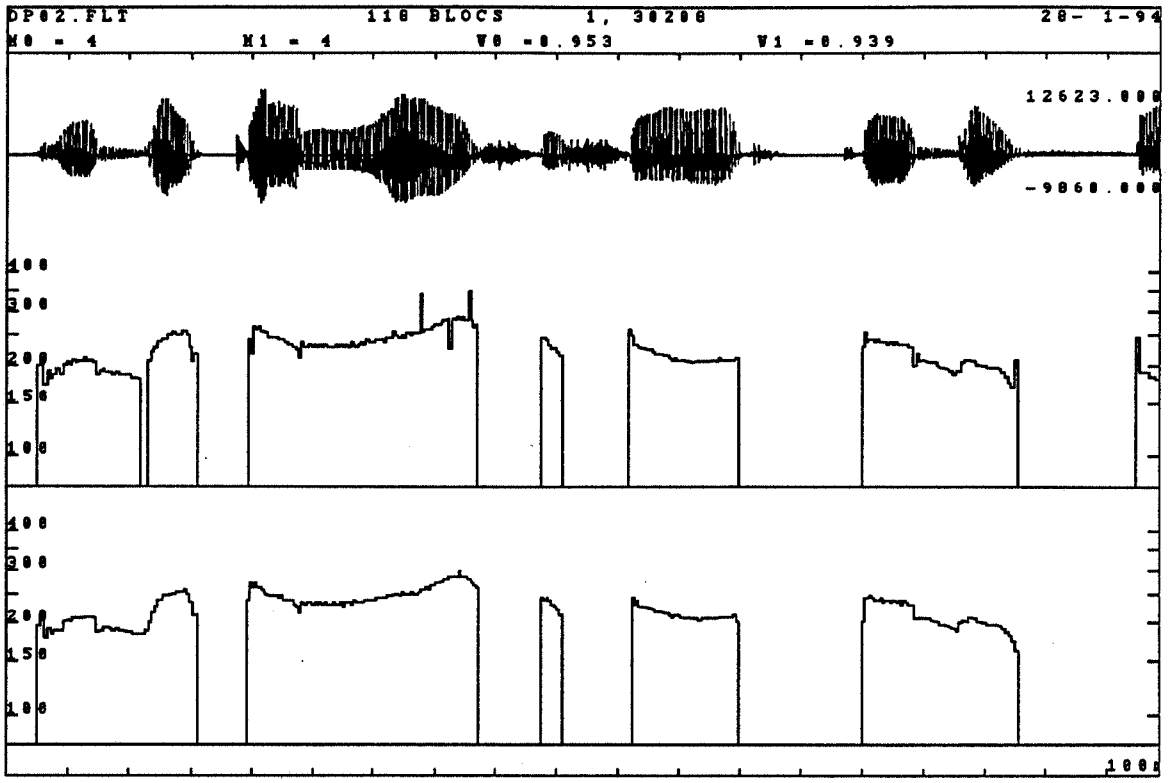


Fig. 7. Résultats de mesure de F_0. De haut en bas : le signal de parole pour la phrase *Les deux camions se sont heurtés de face*, le contour de F_0 en Hz obtenu par notre algorithme, la mesure de référence de F_0 en Hz obtenue par un algorithme classique semi-automatique.

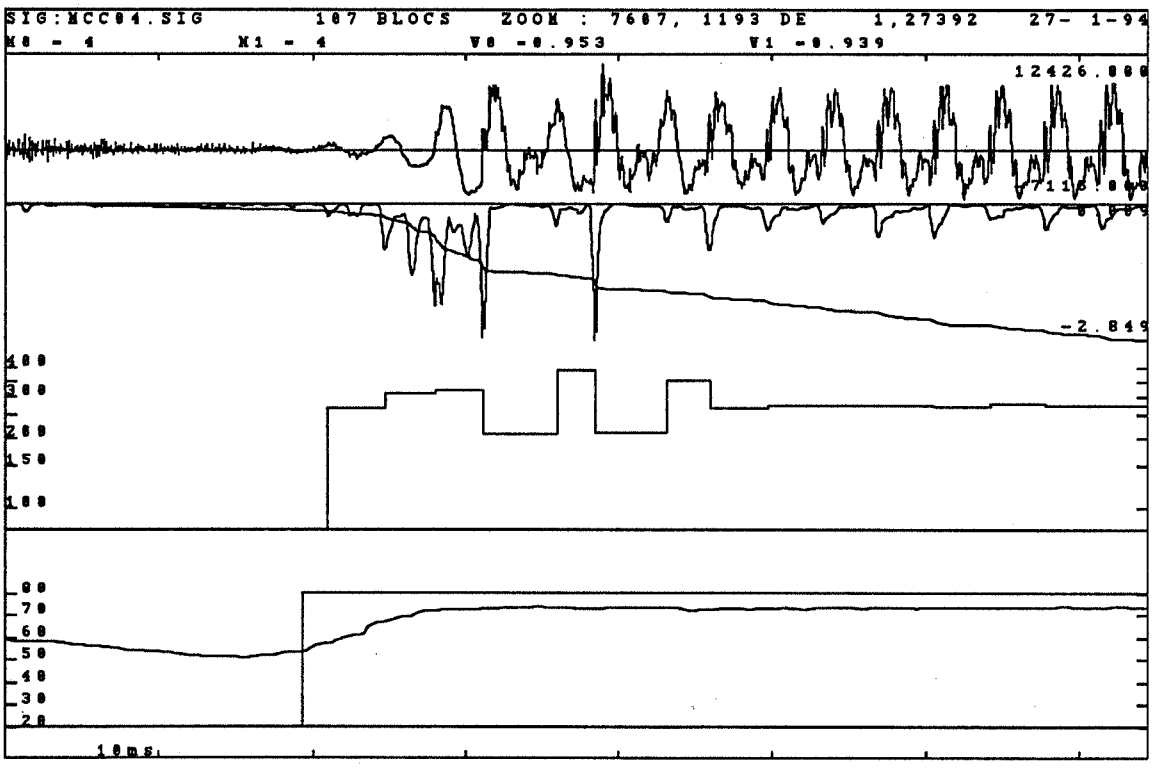


Fig. 8. Résultats de mesure de F_0 sur un segment de *vocal fry*. Même légende que pour la figure 6.

masquent les changements dus à la fermeture de la glotte.

Des segments particulièrement difficiles à traiter sont les segments dits de *vocal fry*, où d'importantes variations énergétiques de la source de voisement donnent un signal particulièrement irrégulier (figure 8). En dépit de ce problème, l'algorithme est capable de localiser correctement la plupart des instants de fermeture de la glotte.

Ces résultats montrent la robustesse de la méthode proposée et confirment les choix des paramètres des deux modèles LPC. Par ailleurs, on peut noter que les paramètres choisis sont valables pour tous les locuteurs.

VI. CONCLUSION

La détection des instants de fermeture de la glotte constitue une méthode précise pour estimer la période fondamentale du signal de parole. La détection automatique de ruptures fournit un outil fiable et performant. Notre expérience montre que l'utilisation de la modélisation adaptative permet d'obtenir de bonnes performances pour tous les sons voisés. Nous avons proposé un nouvel algorithme de détection des instants de fermeture de la glotte pour la mesure de la micromélodie. Un premier test sur un corpus multilocuteur de 24 phrases montre que le système fournit des résultats très prometteurs. La mise au point d'un tel outil ouvre des perspectives intéressantes dans le domaine du filtrage inverse. En effet, la période de fermeture de la glotte provoque un découplage acoustique entre les cavités supra- et subglottiques (Guérin et Boë, 1980). Une détection précise des instants de fermeture permettrait l'extraction de la fonction de transfert du conduit vocal, tâche très difficile pour les méthodes classiques. Une étude de filtrage inverse basée sur notre algorithme de détection est actuellement en cours.

Références

ANANTHAPADMANABHA T. V. & YEGNANARAYANA B. (1979), Epoch extraction from linear prediction residual for identification of closed glottis interval. - I.E.E.E. Trans. ASSP, vol. 27 n° 4, 309-318.

ANDRÉ-OBRECHT R. (1988), A new statistical approach for the automatic segmentation of continuous speech signals. - I.E.E.E. Trans. ASSP, vol. 36 n° 1, 29-40.

ANTONIADES Y. (1993), Étude comparative pour la mesure de la fréquence fondamentale du signal de parole. - Rapport DEA - SIP, INPG.

BASSEVILLE M. & BENVENISTE A. (1983), Sequential detection of abrupt changes in spectral characteristics of digital signals. - I.E.E.E. Trans. Info. Theory, vol. 29 n° 5, 708-723.

BOË L. J. & RAKOTOFIRINGA H. (1971), Exigences, réalisation et limites d'un appareillage destiné à l'étude de l'intensité et de la hauteur d'un signal acoustique. - Acoustique n° 4, 104-113.

CHENG Y. M. & O'SHAUGHNESSY D. (1989), Automatic and reliable estimation of glottal closure instant and period. - I.E.E.E. Trans. ASSP, vol. 38 n° 12, 1806-1814.

DI FRANCESCO R.J. (1990), Real-time speech segmentation using pitch and convexity jump models: application to variable rate speech coding. - I.E.E.E. Trans. ASSP, vol. 38 n° 5, 741-748.

FENG G., ACHAB N. & COMBESCURE P. (1991), On-line speech segmentation using adaptative models: application to variable rate speech coding. - Eurospeech 91, vol. 2, 705-708.

FRIEDLANDER B. (1982), Lattice filters for adaptive processing. - Proceedings of the I.E.E.E., vol. 70, n° 8, 829-867.

GUÉRIN B. & BOË L. J. (1980), Etude de l'influence du couplage acoustique source-conduit vocal sur le F_0 des voyelles orales - conséquences pour l'étude des caractéristiques intrinsèques. - Phonetica, vol. 37, 465-468.

HESS W. (1983), Pitch determination of speech signals. Springer-Verlag, Berlin.

MARKEL J. D. (1972), The SIFT algorithm for fundamental frequency estimation. - I.E.E.E. Trans. Audio Electroacoust., vol. 20, 367-377.

MARTIN P. (1981), Détection de F_0 par intercorrélation avec une fonction peigne. J.E.P., 221-232.

MOULINES E. & DI FRANCESCO R. J. (1990), Detection of glottal closure by jumps in the statistical properties of the speech signal. - Speech Comm., vol. 9, 401-418.

MURGIA C. (1992), Mesure de la fréquence fondamentale du signal de parole à l'aide de la détection automatique de ruptures. - Rapport DEA - SIP, INPG.

NOLL A. M. (1970), Cepstrum pitch determination. - J. Acoust. Soc. Amer., vol. 41 n° 2, 293-309.

ROSS M. J., SHAFFER H. L., COHEN A., FREUDBERG R. & MANLEY H. J. (1974), Average magnitude difference function pitch extractor. - I.E.E.E. Trans. ASSP, vol. 22, 353-362.

SONDHI M. M. (1968), New methods of pitch extraction. - I.E.E.E. Trans. Audio Electroacoust., vol. 16, 262-266.

STRUBE H. W. (1974), Determination of the glottal closure from the speech wave. - J. Acoust. Soc. Amer., vol. 56 n° 5, 1625-1629.

WISE J. D., CAPRIO J. R. & PARKS T. W. (1976), Maximum likelihood pitch estimation. - I.E.E.E. Trans. ASSP, vol. 24, 418-423.

YOUNG T. Y. (1965), Epoch detection - A method for resolving overlapping signals. - Bell Syst. Tech. J., vol. 44, 401-426.

EVALUATION DE LA RECONNAISSANCE DE TRAITS PHONETIQUES EN UTILISANT L'INFORMATION MUTUELLE

Christoph WINDHEUSER(*), Shigeki OKAWA(*)(**),
Frédéric BIMBOT(*), Katsuhiko SHIRAI(**)

(*) Télécom Paris - dépt. Signal, C.N.R.S. - URA 820, Paris, France
(**) Université de Waseda, dépt. Electrical Engineering, Tokyo, Japon

Résumé

Dans cet article nous évaluons les performances d'un perceptron multicouches TDNN appris pour la reconnaissance de traits phonétiques. Pour comparer les différents taux d'erreur, il faut normaliser les taux par les différentes fréquences d'occurrence. Une autre technique pour obtenir une valeur objective de la performance est le calcul de l'information mutuelle entre la sortie du réseau et les sorties désirées. Par nos différentes expériences, nous montrons que les deux mesures sont équivalentes. Les expériences nous donnent les résultats suivants : le réseau est un bon classificateur pour les traits phonétiques avec un taux d'erreur variant entre 1,23 % et 10,27 % sur une base de test indépendante; les traits phonétiques qui décrivent le mode d'articulation sont plus facile à reconnaître par le réseau que les traits qui décrivent le lieu ou le point d'articulation.

Mots-clés : reconnaissance de traits phonétique, perceptron multicouche TDNN, information mutuelle.

1. INTRODUCTION

Une approche classique pour la reconnaissance de la parole consiste à utiliser un *système hybride*. Un tel système se compose d'un système connexionniste pour la reconnaissance statique d'une représentation intermédiaire et une procédure d'alignement temporel. Un grand avantage d'un système connexionniste est sa capacité à apprendre. Il existe des procédures d'apprentissage pour optimiser ses paramètres. Malheureusement, pour un problème réel comme la reconnaissance de la parole, on a besoin d'un grand nombre de paramètres, ce qui exige normalement une base de données énorme pour l'apprentissage. Une solution peut consister à utiliser le savoir d'un expert pour mieux adapter le système au problème et réduire le nombre de paramètres.

En général, la représentation intermédiaire d'un système hybride utilise les phonèmes avec une unité de sortie pour chacun d'eux. Mais il est possible de changer cette représentation canonique en une représentation plus compacte, par exemple une représentation par *traits phonétiques*. Cette représentation permet une description des phonèmes avec moins de paramètres.

Dans un article précédent [5], nous avons montré les très bons résultats obtenus avec un système hybride utilisant des traits phonétiques pour la reconnaissance globale de mots. Dans cet article nous étudions plus précisément les performances pour chacun des traits. Pour mieux les comparer, nous évaluons *l'information mutuelle* entre l'activation de la sortie observée et celle de la sortie désirée.

2. LA BASE DE DONNEES

Pour évaluer notre système connexionniste, nous avons choisi la base de données ALPH de l'Université Carnegie Mellon aux Etats-Unis. C'est une base monoloculeur de lettres épelées de l'alphabet anglais. Elle comprend 1000 séquences de lettres et chaque séquence comprend 7 à 9 lettres épelées. La parole enregistrée est de haute qualité. Elle est analysée avec un processeur acoustique pour obtenir 16 coefficients de bancs de filtre toutes les 10 ms. Dans les expériences que nous allons décrire ici, nous avons séparé la base de données en 500 séquences pour l'apprentissage et 500 séquences pour le test.

3. LE SYSTEME

Pour la reconnaissance de traits phonétiques, nous avons choisi un TDNN *(Time Delay Neural Network)* proposé par Waibel et al. [4]. Un TDNN est un perceptron multicouches avec délais temporels. Les poids du TDNN sont appris par rétropropagation du gradient. La figure 1

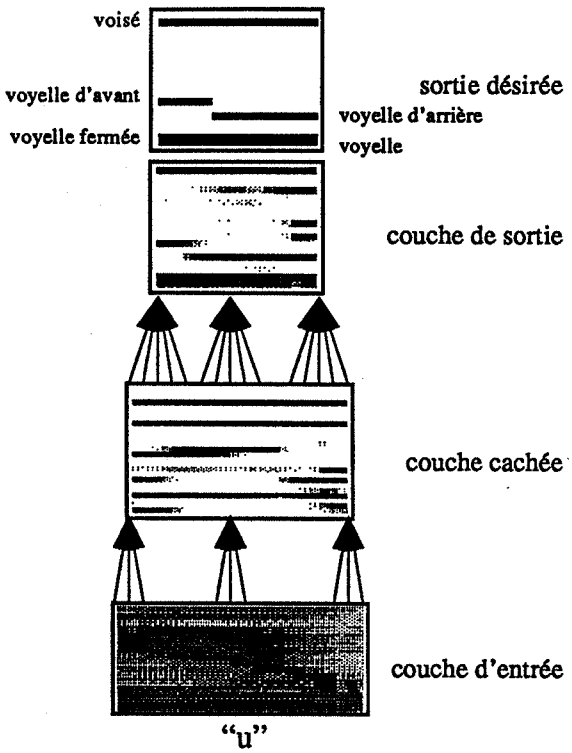


Figure 1 : Le perceptron multicouches utilisé pour la reconnaissance de traits phonétiques.

présente une vue d'ensemble du système.

Le système se compose de trois couches : la couche d'entrée, la couche cachée et la couche de sortie. La couche d'entrée a 16 unités correspondant aux 16 coefficients d'entrée, la couche cachée a 24 unités et la couche de sortie a une unité pour chaque trait phonétique. Les connexions entre la couche d'entrée et la couche cachée ont des délais temporels de -1, 0 et 1. Les connexions entre la couche cachée et la couche de sortie ont des délais temporels de -2, -1, 0, 1 et 2.

Pour l'apprentissage, nous utilisons la rétropropagation du gradient "on-line" pour chaque trame sans momentum et un taux d'apprentissage de 1,0. Les sorties désirées pour l'apprentissage sont fournies par un système d'étiquetage automatique (par Modèles de Markov). Après 150 présentations des données d'apprentissage, les résultats sont stables et nous arrêtons l'apprentissage.

4. L'ENSEMBLE DES TRAITS PHONETIQUES

La définition d'un ensemble de traits phonétiques n'est pas simple. Il en existe un grand nombre dans la littérature. Pour la reconnaissance de la parole, des critères importants pour bien choisir un ensemble de traits sont :

- L'ensemble choisi doit permettre de distinguer tous les phonèmes, donc les traits doivent être discriminants.
- Les traits doivent être bien reconnus par le système connexionniste.

Nous avons fait de nombreuses expériences avec les différents ensembles décrits dans [5] et nous avons obtenu les meilleurs résultats avec un ensemble qui se compose de 18 traits. Outre les 2 traits *voyelle* et *consonne*, on utilise 6 traits pour les voyelles *(voyelle fermée, voyelle mi-ouverte, voyelle ouverte, voyelle d'arrière, voyelle centrale* et *voyelle d'avant)* et 8 traits pour les consonnes *(consonne vocalique, liquide, nasale, plosive, fricative, consonne d'avant, consonne centrale* et *consonne d'arrière)*. De plus, on adjoint 2 traits complémentaires: *voisé* et *non voisé*. Une propriété de notre système de traits est que chaque phonème est représenté par 4 traits actifs (égaux à 1) exactement. La table 1 donne la définition exacte des traits pour chaque phonème.

Nous rappelons ici les résultats en reconnaissance globale de mots avec ces traits phonétiques : pour 26 mots correspondant aux 26 lettres de l'alphabet anglais, nous avons obtenu 96,1 % de reconnaissance, contre 93,8 % avec une représentation phonétique canonique [5], sur la base ALPH décrite au paragraphe 2.

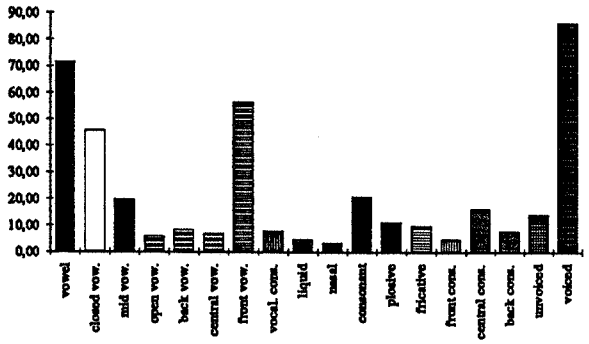


Figure 2 : Pourcentage de trames, pour lesquelles un trait phonétique est actif.

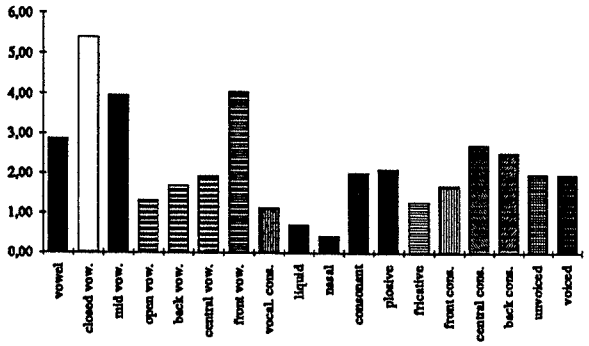


Figure 3 : Erreur totale pour chaque trait phonétique.

	aa	ah	ax	eh	y	iy	ow	uw	l	m	n	jh	r	z	v	w	d	b	s	ch	f	t	k	p
voyelle	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
voyelle fermée	-	-	-	-	+	+	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
voyelle mi-ouverte	-	+	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
voyelle ouverte	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
voyelle d'arrière	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
voyelle centrale	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
voyelle d'avant	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
consonne vocalique	-	-	-	-	-	-	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-
liquide	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
nasale	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
consonne	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	+	+	+	+	+	+	+	+
plosive	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	+	+
fricative	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-	-	+	+	+	-	-	-
consonne d'avant	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	+	-	-	+	-	-	+
consonne centrale	-	-	-	-	-	-	-	-	+	-	+	-	-	+	-	-	+	-	+	-	-	+	-	-
cons. d'arrière	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	+	-	-	+	-
non voisé	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
voisé	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-

Table 1 : Définition de l'ensemble des traits phonétiques.

5. RECONNAISSANCE DE TRAITS PHONETIQUES

Les scores de reconnaissance de mots ne donnent pas d'information directe sur les traits bien reconnus et ceux mal reconnus. Dans cet article, nous évaluons la performance des traits individuellement.

Parce que les valeurs d'activation des unités de sortie sont continues entre 0 et1, une solution simple est l'utilisation d'un seuil. Quand l'activation d'une unité est plus grande que le seuil, on considére que le trait correspondant est reconnu par le réseau. Nous avons choisi une valeur de 0,5 comme seuil. Ce seuillage est effectué pour chaque trame.

Pour évaluer le taux de reconnaissance des traits, nous comparons les traits reconnus par le réseau pour chaque trame avec les sorties desirées pour ces mêmes trames. Cette évaluation est réalisée sur une base de test indépendante de la base d'apprentissage. Ces résultats sont donnés dans la deuxième colonne de la table 2 et sont repris sous forme graphique dans la figure 3. Comme cela a été montré dans [1], cette *erreur totale* est dépendante des fréquences d'occurrence. On peut illustrer cela par un exemple trivial :

Imaginons une unité de sortie non apprise qui donne l'activation zéro tout le temps. Si un trait a une fréquence d'occurrence de 10 %, cette unité fait une erreur totale de 10 %. Mais si la fréquence d'occurrence est de 50 %, l'erreur totale de la même unité est de 50 %. Cet exemple confirme qu'il est difficile de comparer les résultats pour différentes fréquences d'occurrence.

La première colonne de la table 2 et la figure 2 donnent donc les fréquences d'occurrence. On observe qu'il existe de grandes différences entre les traits. Le trait *voisé* par exemple apparaît 26 fois plus souvent que le trait *nasale*. Si l'on compare l'erreur totale (figure 3) avec les fréquences d'occurrence (figure 2), il est manifeste que les traits ayant une faible fréquence d'occurrence comme les traits *voyelle ouverte, voyelle d'arrière, voyelle centrale, consonne vocalique, liquide, nasale* et *fricative* ont aussi une petite erreur totale. Donc, cette mesure d'erreur n'est pas très significative.

En fait, l'erreur totale est une somme de deux erreurs. La première, quand la sortie desirée est active et la sortie réelle inactive; la deuxième, quand la sortie desirée n'est pas active et la sortie réelle active. On peut appeler la première erreur *erreur positive* et la deuxième *erreur negative*. Les colonnes 3 et 4 de la table 2 ainsi que les figures 4 et 5 donnent ces deux taux d'erreur. On voit que ces taux peuvent être très differents de l'erreur totale. Par exemple, le trait *consonne d'avant* a une erreur totale relativement faible valant 1,67 %. Mais si l'on regarde l'erreur positive, on voit que pour plus de 20 % des trames où le trait *consonne d'avant* est actif, le réseau ne le reconnaît pas correctement. Dans ce cas, la performance du réseau est plus mauvaise que l'erreur totale ne le suggère. L'erreur négative est généralement plus basse que l'erreur positive parce que, dans notre base de données, il existe beaucoup plus de sorties désirées négatives que positives.

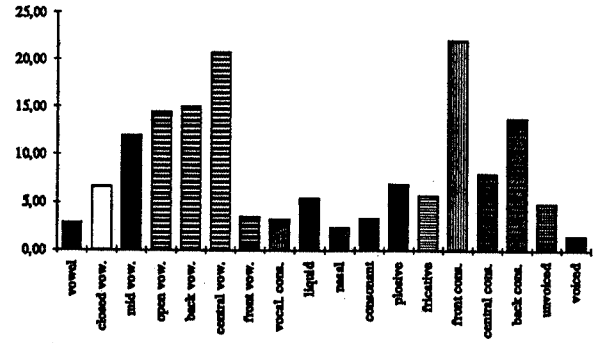


Figure 4 : Erreur positive sur la base de test.

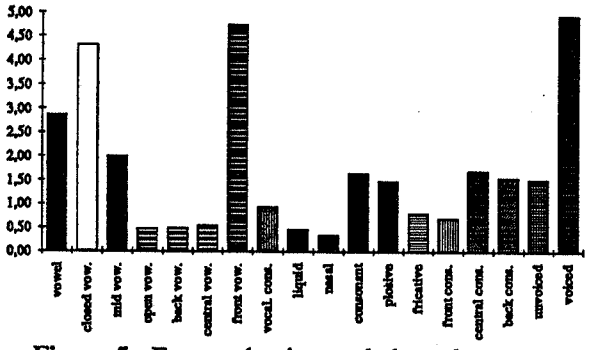


Figure 5 : Erreur négative sur la base de test.

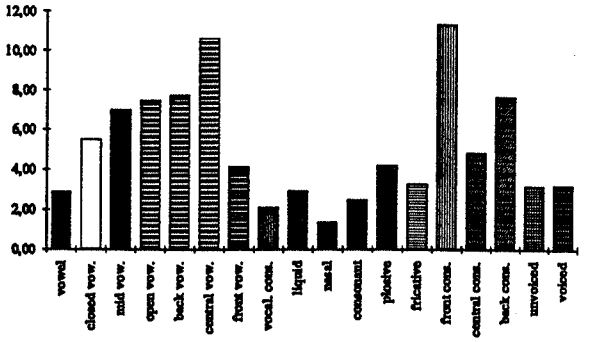


Figure 6 : Erreur moyenne pour tous les traits phonétique sur la base de test.

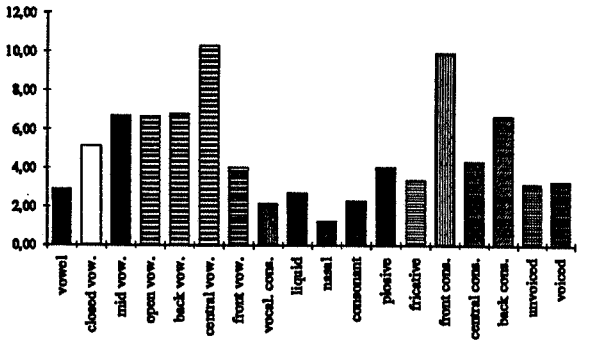


Figure 7 : Erreur moyenne calculée avec la fonction *maximum* sur la base de test.

Pour obtenir une valeur qui représente la performance de la reconnaissance pour chaque trait, nous calculons la moyenne de l'erreur positive et de l'erreur négative. Nous appellons cette erreur *l'erreur moyenne*. Elle est donnée dans la colonne 5 de la table 2 ainsi que dans la figure 6. Ces valeurs nous permettent de donner une première interprétation de la performance de notre réseau :

On voit que tous les traits *voyelle fermée, voyelle mi-ouverte, voyelle ouverte, voyelle d'arrière, voyelle centrale, consonne d'avant* et *consonne d'arrière* ont une erreur moyenne plus grande que les traits *voyelle, consonne vocalique, liquide, nasal* et *consonne*, ainsi que *voisé, non voisé* et *plosive, fricative*. On peut dire que les traits phonétiques relatifs au mode d'articulation sont mieux reconnus par le réseau que les traits phonétiques relatifs au lieu d'articulation.

6. CLASSIFICATION BAYESIENNE

Comme il est démontré par exemple dans [2], il est possible de considérer le système connexionniste appris par minimisation de l'erreur quadratique moyenne comme l'approximation d'un *classificateur bayésien*. Cela signifie que les activations des unités de sortie estiment la probabilité conditionnelle que la sortie désirée soit 1 pour la trame d'entrée considérée (probabilité a posteriori). Pour notre système de 18 traits, nous pouvons exploiter le fait que chaque phonème correspond à exactement 4 traits actifs : plutôt que d'utiliser un seuillage fixe, nous sélectionnons les 4 sorties d'activité maximale (fonction maximum). Comme on peut le voir dans la colonne 6 de la table 2 et sur la figure 7, les résultats sont légèrement meilleurs avec la fonction maximum qu'avec le seuil, mais les relations entre les différents traits ne changent pas.

7. EVALUATION DE L'INFORMATION MUTUELLE

Nous voulons étudier d'autres possibilités pour évaluer la performance de notre système. Bredshaw et Bell [1] ont proposé une mesure connue dans le domaine de la détection de signaux, sans en donner une définition explicite. Nous utilisons une mesure plus générale et mieux connue, *l'information mutuelle*, qui a été introduite par Shannon [3].

Avec le calcul de l'information mutuelle entre deux événements aléatoires, on peut mesurer l'information contenue dans un événement par rapport à un autre. On peut considérer que le but de l'apprentissage d'un système connexionniste est de maximiser l'information dans la couche de sortie relativement aux sorties desirées. Donc l'information mutuelle entre la sortie observée et la sortie désirée peut être une mesure objective de la qualité d'apprentissage.

La définition de l'information mutuelle $I(X;Y)$ entre

		% T.A.	% E.T.	% E.P.	% E.N.	% E.M.	%E.M.M.
voyelle	(vowel)	71,34	2,88	2,89	2,87	2,88	2,91
voyelle fermée	(closed vowel)	45,69	5,39	6,67	4,31	5,49	5,12
voyelle mi-ouverte	(mid vowel)	19,62	3,95	11,96	2,00	6,98	6,70
voyelle ouverte	(open vowel)	6,03	1,31	14,41	0,47	7,44	6,66
voyelle d'arrière	back vowel)	8,28	1,69	14,93	0,49	7,71	6,80
voyelle centrale	(central vowel)	6,85	1,91	20,62	0,53	10,58	10,27
voyelle d'avant	(front vowel)	56,21	4,04	3,49	4,74	4,11	4,02
consonne vocalique	(vocalic consonant)	7,81	1,12	3,22	0,94	2,08	2,17
liquide	(liquid)	4,61	0,69	5,38	0,46	2,92	2,72
nasale	(nasal)	3,20	0,41	2,37	0,34	1,35	1,23
consonne	(consonant)	20,78	1,99	3,29	1,65	2,47	2,26
plosive	(plosive)	11,22	2,08	6,89	1,48	4,18	4,01
fricative	(fricative)	9,56	1,27	5,68	0,80	3,24	3,37
consonne d'avant	(front consonant)	4,61	1,67	21,93	0,69	11,31	9,88
consonne centrale	(central consonant)	16,13	2,71	7,98	1,70	4,84	4,31
cons. d'arrière	(back consonant)	7,85	2,50	13,73	1,54	7,64	6,65
non voisé	(unvoiced)	13,91	1,97	4,86	1,51	3,18	3,16
voisé	(voiced)	86,02	1,94	1,46	4,93	3,20	3,31
moyenne		22,21	2,20	8,43	1,75	5,09	4,75

Table 2 : Résultats des expériences décrites dans les paragraphes 5 et 6.
T.A. = traits phonétiques actifs, E.T. = erreur totale, E.P. = erreur positive, E.N. = erreur négative, E.M. = erreur moyenne, E.M.M. = erreur moyenne calculée avec la fonction maximum.

deux événements aléatoires X et Y se base sur *l'entropie* $H(X)$:

$$H(X) = -\sum_{x} P(X=x) \log P(X=x) \quad (1)$$

De la même façon, *l'entropie conditionnelle*, qui mesure l'incertitude d'un événement aléatoire X, connaissant le résultat d'un événement aléatoire Y, est définie par :

$$H(X|Y) = -\sum_{x,y} P(X=x,Y=y) \log P(X=x|Y=y) \quad (2)$$

L'information mutuelle est définie par :

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} P(X=x,Y=y) \log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} \quad (3)$$

La valeur $I(X;Y)$ est toujours dans l'intervalle $[0; H(X)]$. Pour comparer l'information mutuelle entre les différentes unités de sortie, nous utilisons *l'information mutuelle normalisée* $N(X;Y)$:

$$N(X;Y) = \frac{I(X;Y)}{H(X)} \quad (4)$$

Pour calculer l'information mutuelle normalisée entre les sorties du réseau et les sorties désirées, il faut estimer les distributions de probabilités avec un histogramme.

La table 3 donne l'entropie des unités de sortie $H(O)$, l'entropie des sorties désirées $H(T)$, l'information mutuelle $I(T;O)$ et l'information mutuelle normalisée $N(T;O)$. La figure 8 donne les valeurs $1 - N(T;O)$ pour la base de test. Comme on le constate, il existe de nombreuses concordances entre cette mesure et la mesure d'erreur moyenne (figure 6 et 7).

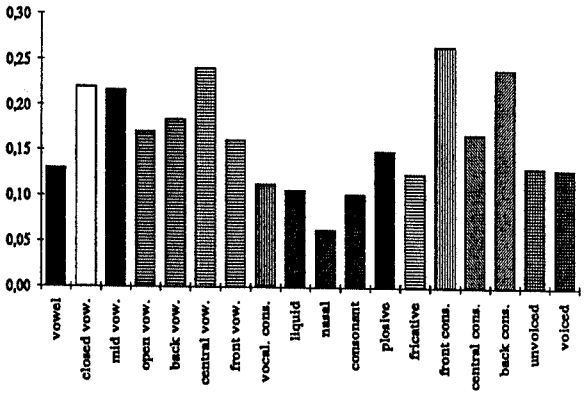


Figure 8 : 1 - N(T;O) pour tous les traits phonétique sur la base de test, où N(T;O) est l'information mutuelle normalisée.

	H(O)	H(T)	I(T;O)	N(T;O)
voyelle	2,40	0,86	0,75	0,87
voyelle fermée	3,23	0,99	0,78	0,78
voyelle mi-ouverte	2,50	0,71	0,56	0,78
voyelle ouverte	1,35	0,33	0,27	0,83
voyelle d'arrière	1,55	0,41	0,34	0,82
voyelle centrale	1,62	0,36	0,27	0,76
voyelle d'avant	3,28	0,99	0,83	0,84
cons. vocalique	1,18	0,40	0,35	0,89
liquide	0,79	0,27	0,24	0,89
nasale	0,72	0,20	0,19	0,94
consonne	1,96	0,74	0,66	0,90
plosive	1,58	0,51	0,43	0,85
fricative	1,44	0,45	0,40	0,88
consonne d'avant	1,14	0,27	0,20	0,74
consonne centrale	2,20	0,64	0,53	0,83
cons. d'arrière	1,58	0,40	0,30	0,76
non voisé	1,56	0,58	0,51	0,87
voisé	1,57	0,58	0,51	0,87
moyenne	1,76	0,54	0,45	0,84

Table 3 : Entropie des unités de sortie et des sorties désirées, information mutuelle et information mutuelle normalisée sur la base de test.

De nouveau, tous les traits phonétiques qui décrivent le lieu d'articulation, comme *voyelle fermée, voyelle mi-ouverte, voyelle ouverte, voyelle d'arrière, voyelle centrale, voyelle d'avant, consonne d'avant, consonne centrale* et *consonne d'arrière*, ont une information mutuelle normalisée qui est inférieure ou égale à la moyenne. Tous les autres traits phonétiques, qui décrivent le mode d'articulation, ont une information mutuelle normalisée qui est plus grande que la moyenne.

Cela nous conduit à la réflexion suivante : l'information mutuelle normalisée entre l'activation des unités de sortie et les sorties désirées est une mesure objective de la performance du réseau qui est cohérente avec les taux d'erreur. La figure 9 représente graphiquement la relation qui existe entre les deux quantités.

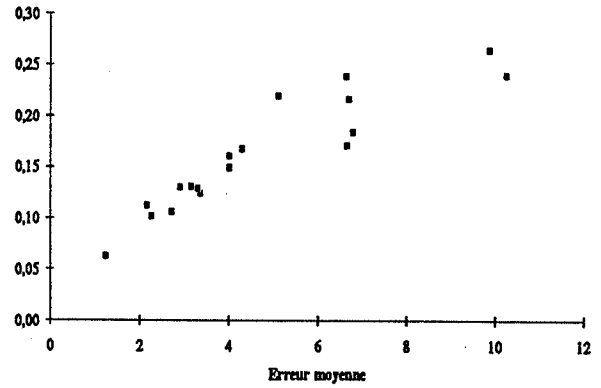


Figure 9 : Information mutuelle normalisée en fonction de l'erreur moyenne calculée avec la fonction *maximum*.

8. CONCLUSIONS

Dans cet article, nous avons évalué les taux d'erreur d'un perceptron multicouches TDNN, pour la classification de traits phonétiques. Pour comparer les différents taux, nous avons montré qu'il était important de tenir compte des différentes fréquences d'occurrence. Nous avons présenté une autre technique pour évaluer la performance du réseau : le calcul de l'information mutuelle entre l'activation de la sortie et l'activation de la sortie désirée.

Les résultats, qui sont cohérents avec les taux d'erreur normalisés par les différentes fréquences d'occurrence, nous permettent de dire :

- Un réseau appris pour la reconnaissance de traits phonétique donne une bonne performance avec des taux d'erreur moyenne entre 1,23 % et 10,27 % pour la base de test.
- La reconnaissance des traits phonétiques qui décrivent le mode d'articulation est plus facile que la reconnaissance des traits qui décrivent le lieu ou le point d'articulation.
- Il existe une très forte corrélation entre le taux d'erreur moyenne et l'information mutuelle normalisée.

BIBLIOGRAPHIE

[1] G. Bradshaw and A. Bell. "Towards the Performance Limits of Connectionist Feature Detectors". In *Proc. of the ICSLP*, Banff, Canada, 1992.

[2] M. D. Richard and R. P. Lippmann. "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities". *Neural Computation*, 3, pp. 461-483, 1991.

[3] C. E. Shannon. "Mathematical Theory of Communication". Bell System Technical Journal, 1948.

[4] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang. "Phoneme Recognition using Time-Delay Neural Networks". In *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 328-339, 1989.

[5] C. Windheuser and F. Bimbot. "Phonetic Features for Spelled Letter Recognition with a Time Delay Neural Network", In *Proc. of Eurospeech*, pp. 1489-1492, Berlin, 1993.

MODELES STOCHASTIQUES D'ORDRES 1 et 2

Jean-François MARI Dominique FOHR Jean-Paul HATON

CRIN-CNRS & INRIA Lorraine

BP 239

54506 VANDOEUVRE-LES-NANCY

Résumé

En reconnaissance automatique de la parole à l'aide de modèles stochastiques, on suppose qu'un mot est constitué par une suite de segments stationnaires et indépendants. Cet article présente deux méthodes pour mieux modéliser la durée de ces segments. Une première approche consiste à utiliser des modèles stochastiques d'ordre deux, l'autre consiste à modéliser la suite des durées relatives des segments d'un mot à l'aide de méthodes d'analyse de données. Après une définition mathématique des modèles de Markov d'ordre deux et une présentation de notre modèle de durée, nous donnerons les résultats sur la base TIDIGITS où une réduction de 10% du taux d'erreur est observée.

Mots-clés: modèles de Markov continus d'ordre 2, reconnaissance de chiffres enchaînés, durées.

1. INTRODUCTION

Au cours de ces quinze dernières années, un travail intensif en reconnaissance de la parole a été réalisé en utilisant des modèles de Markov du premier ordre (HMM1). Toutes ces études supposent que la parole est constituée d'une suite de segments indépendants et stationnaires. Dans cet article, nous commencerons par conduire une étude comparative de ces modèles du premier ordre avec des modèles du second ordre (HMM2). Après avoir donné les définitions mathématiques des HMM2, nous donnerons les détails d'implantation pour réduire leur complexité algorithmique. La reconnaissance se fait à l'aide de l'algorithme de recherche des N meilleurs chemins ce qui permet de prendre en compte d'autres sources de connaissance pour améliorer la reconnaissance. Dans ce cadre, nous proposerons l'ajout d'un modèle de durée pour éliminer les solutions présentant des alignements incorrects entre la suite des états cachés d'un HMM2 et l'ensemble des observations.

2. MODELES DE MARKOV DU 2^e ORDRE

Dans un modèle du second ordre (HMM2) la suite d'états cachés est supposée être une suite de Markov du second ordre, c'est-à-dire que la probabilité de transition entre états dépend des états dans lesquels le modèle se trouvait aux temps t-1 et t-2.

2.1. Notations

On appellera:

- λ le modèle de Markov d'ordre 2,
- S_t la variable aléatoire au temps t dont les réalisations sont les états du modèle,
- b_i la densité associée à l'état i.
- O_t l'observation au temps t (de dimension D),
- $P(O/\lambda)$ la vraisemblance de la suite d'observations $O_1, O_2 \ldots O_T$ avec le modèle λ
- $\aleph(\mu, \Sigma)$ une loi normale de dimension D de moyenne μ et de matrice de covariance Σ
- on notera M^t la transposée de la matrice M.

2.2. Passage de l'ordre 1 à l'ordre 2

Classiquement, les probabilités de transition entre états d'un HMM1 sont :

$$\text{Prob}(S_t = k \mid_{S_{t-1}=j,\, S_{t-2}=i,\, S_{t-3}=\ldots}) = \text{Prob}(S_t = k \mid_{S_{t-1}=j}) = a_{jk}$$

Dans un HMM2 elles deviennent :

$$\text{Prob}(S_t = k \mid_{S_{t-1}=j,\, S_{t-2}=i,\, S_{t-3}=\ldots}) = \text{Prob}(S_t = k \mid_{S_{t-1}=j,\, S_{t-2}=i}) = a_{ijk}$$

Les densités associées aux états restent multinormales

$$b_i(O_t) = \sum_m c_{im}\, \aleph(\mu_m, \Sigma_m, O_t) \qquad \sum_m c_{im} = 1$$

$$\aleph(\mu, \Sigma, x) = \frac{1}{\sqrt{(2\pi)^D \det \Sigma}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}$$

Les probabilités "Forward-Backward" s'obtiennent simplement en ajoutant un indice indiquant l'état d'où on vient :

$$\alpha_{t+1}(j,k) = \sum_{i=1}^{S} \alpha_t(i,j) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

$$\beta_t(i,j) = \sum_{k=1}^{S} \beta_{t+1}(j,k) a_{ijk} b_k(O_{t+1}), \quad 2 \le t \le T-1$$

les compteurs de transitions deviennent:

$$\eta_t(i,j,k) = \frac{\alpha_t(j,k) a_{ijk} b_k(O_{t+1}) \beta_{t+1}(j,k)}{P(O/\lambda)}, \quad 1 \le t \le T-1$$

A partir de ces définitions, la réestimation, suivant le maximum de vraisemblance, des paramètres des modèles est classique [Kriouile90]

3. COMPLEXITE

Dans le cas d'un HMM1, la complexité théorique est en SxSxT, S étant le nombre d'états et T le nombre de observations. En pratique, la sommation dans les calculs de α et β n'est faite que pour les a_{ij} non nuls; ce qui ramène la complexité en Ne x S x T, où Ne est le nombre moyen de transitions entrantes pour l'ensemble des états.

Dans le cas d'un HMM2, la complexité théorique devient SxSxSxT. Comme pour les HMM1, la complexité pratique se ramène à Ne x Ne x S x T.

La figure 1. montre la topologie du modèle utilisée dans cette étude. Le nombre moyen de transitions entrantes est donc de 2, ce qui conduit à un ensemble d'algorithmes pour l'apprentissage et pour le test deux fois plus lents, ce qui reste raisonnable.

Figure 1: topologie des modèles utilisés.

Pour des facilités d'initialisation, les deux premiers états ne possèdent pas de densité d'émission.

4. RECONNAISSANCE

L'algorithme de Viterbi se généralise sans difficulté au deuxième ordre pour un modèle considéré. En revanche, les suites de mots dans le réseau linguistique obéissent à un processus de Markov d'ordre 1. Afin de pouvoir, par la suite, utiliser d'autres traitements ou sources de connaissances, nous recherchons les N meilleurs chemins tel que décrit dans [Schwartz91].

5. MODELES DE DUREE

La durée d'un état correspond à la durée d'un segment stationnaire où toutes les trames sont indépendantes les unes par rapport aux autres. Dans un HMM1, cette durée suit une loi géométrique. Il a été noté que cette hypothèse ne correspond pas à la réalité. Différents travaux ont porté sur la durée des états d'un HMM1 [Levinson86] [Wilpon93] à l'aide de densités paramétriques ou non. Ils ont conduit à la définition de semi-modèles de Markov. D'autres auteurs ont modélisé la corrélation entre trames à l'intérieur d'un segment [Wellekens87] [Digalakis91] [Saerens93]. Dans un HMM2, la durée d'un état ne suit pas une loi géométrique simple. Un HMM2 suppose aussi l'indépendance des trames entre elles à l'intérieur d'un segment. Nous avons noté que le modèle HMM2 converge rapidement vers un modèle où, pour certains états, notés j par la suite, la probabilité de transition a_{ijk} telque $i \neq j \neq k \neq i$ devient rapidement nulle, ce qui signifie que le processus de Markov doit rester pendant plus d'une fois dans l'état j. Ce résultat intéressant nous fait penser au modèle de Ferguson qui a été développé pour les HMM1. En effet si on représente le modèle HMM1 engendrant la même suite d'états cachés qu'un HMM2 on trouve que les états dont la durée est supérieure à un sont les états dupliqués du modèle de Ferguson. Il est intéressant de noter qu'un HMM2 converge naturellement vers le modèle de Ferguson.

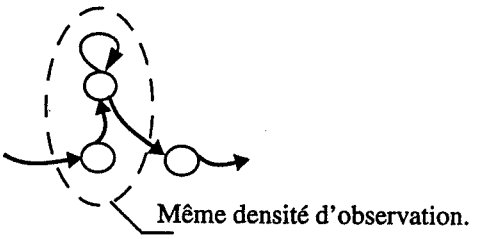


Figure 2: état pour lequel a_{ijk} est nul.

L'étude des erreurs produites par l'algorithme de Viterbi pour le second ordre montre qu'il est nécessaire d'avoir un mécanisme pour rejeter certaines réponses correspondant à des alignements singuliers. Nous avons choisi d'utiliser la durée des segments pour rejeter ces alignements. Une façon classique de modéliser la durée d'un état est de considérer que le nombre de trames capturées par un état suit une loi de probabilité d'une famille donnée *a priori*. Différentes familles ont été utilisées : normale, gamma, Poisson, [Suaudeau93]. Nous avons effectué une reconnaissance sur le corpus d'apprentissage afin d'avoir une statistique sur le nombre de trames capturées par les états de chaque HMM2. Cette étude a montré une forte corrélation des durées entre états. Pour tenir compte de cette corrélation, nous

avons préféré effectuer une classification des durées au niveau du mot en regroupant, dans une classe par mot, tous les alignements corrects.

Un alignement est obtenu à l'aide de l'algorithme de Viterbi. Il est représenté par un vecteur constitué des durées relatives de chaque état. Comme la somme des composantes est égale à un, on remplace une des composantes par le nombre total de trames alignées sur le modèle, c'est à dire la durée absolue du mot. Pour chaque classe, correspondant à chacun des mots du vocabulaire, on estime une moyenne et une matrice de covariance. On mesure la distance d'un alignement fourni par l'algorithme de Viterbi à sa classe à l'aide de la métrique de Mahalanobis.

$$d^2(x, g_m) = (x - g_m)^t V_m^{-1} (x - g_m)$$

V_m représente la matrice de covariance du mot m,

g_m représente le vecteur moyen du mot m.

Cette distance vient pondérer la probabilité d'alignement donné par l'algorithme de Viterbi: le score final est une combinaison linéaire de cette probabilité et de la distance de Mahalanobis.

$$finalscore = \alpha \cdot d(x, g_m) + \beta \cdot \log P(O/\lambda_m)$$

Ce score final permet de trier les cent meilleures réponses fournies par l'algorithme de reconnaissance..

6. RESULTATS EXPERIMENTAUX

6.1. Bases de données

Dans cette étude, nous avons utilisé deux corpora, TIDIGITS [Leonard84] et BDSONS [Carré84].

TIDIGITS contient des chiffres enchaînés prononcés par 225 locuteurs Américains (112 pour l'apprentissage, 113 pour le test). Le vocabulaire est constitué de 11 mots ("zero", "one",..."nine", "oh"). Chaque locuteur a prononcé 77 séquences de 1, 2, 3, 4, 5 et 7 chiffres.

BDSONS est la base enregistrée par le GDR-PRC Communication Homme-Machine. Nous avons utilisé la partie concernant les lettres isolées (alr) et connectées (rec). Nous avons choisi 22 locuteurs pour l'apprentissage et 8 autres pour le test.

6.2. Paramétrisation

Nous avons paramétré les deux corpora à l'aide de MFCC (Mel Frequency Cepstrum Coefficient). Les 12 coefficients sont calculés sur une trame de 512 points; le décalage entre deux trames est de 256 points. Le premier coefficient (loudness) n'a pas été utilisé dans les expériences sur TIDIGITS. Pour cette base, les coefficients de régression du premier et du second ordre (notés Δ et $\Delta\Delta$) ont été ajoutés aux 11 coefficients statiques pour chaque trame. En revanche, le nombre de locuteurs de BDSONS n'était pas suffisant pour estimer correctement les paramètres des HMM utilisant des coefficients dynamiques.

6.3. Comparaison entre les modèles d'ordre 1 et 2.

Les HMM1 et HMM2 sont des modèles continus. Ils possèdent tous la même topologie (cf. figure 1.). Sauf pour la pause qui possède 2 états, tous les modèles sont composés de 6 états. Les densités associées à chacun des états sont multi-normales avec des matrices de covariances triangulaires (*full covariance matrices*).

La comparaison a porté sur les 4311 suites de chiffres prononcés par les locuteurs masculins de TIDIGITS. Le système autorise des suites de longueur quelconque. Le tableau 1 contient le taux d'erreur de chaînes pour chacune des méthodes ainsi que les intervalles de confiance au risque de 5%.

On peut remarquer que les performances du modèle du deuxième ordre sont toujours meilleures que celle obtenues à partir du premier ordre, quelque soit la paramétrisation. et cela de façon statistiquement significative.

Tableau 1 : Comparaison HMM1 HMM2
taux d'erreur de chaînes

paramétrisation	HMM1	HMM2
11 MFCC+11Δ+ΔE+ΔΔE	3.6% [3.0%-4.2%]	2.5% [2.1%-3.0%]
11 MFCC+11Δ+ΔE+ 11ΔΔ + ΔΔE	3.3% [2.8%-3.9%]	2.1% [1.7%-2.5%]

6.4. Apport des modèles de durée

Dans le but d'évaluer l'apport de notre modèle de durée, nous avons testé le système sur la totalité des locuteurs hommes et femmes de TIDIGITS. Nous avons utilisé un modèle de silence et deux modèles par chiffre, un pour les hommes et un pour les femmes. Afin de tenir compte de la différence de vitesse d'élocution entre des chiffres isolés et des séquences de chiffres, deux estimations de durée ont été effectuées pour chaque modèle, l'une pour les mots isolés, l'autre pour les séquences de chiffres. Le choix du modèle du durée se fait en utilisant le nombre de mots dans la meilleure réponse de l'algorithme de reconnaissance (N-meilleurs). Le tableau 2 montre que le modèle de durée permet de réduire le taux d'erreur de 10% pour les deux paramétrisations étudiées. On peut noter au passage le faible apport (non statistiquement significatif) des coefficients de régression d'ordre 2 ce qui corrobore les résultats de Brian Hanson [Hanson90]). Dans cette comparaison, les intervalles de confiance (au risque de 5%) se recouvre et il est néces-

saire d'étudier la répartition des erreurs. On trouvera dans [Mari94] une validation statistique de ces résultats..

Tableau 2 : taux d'erreur de chaînes sur TIDIGITS

Paramétrisation	sans durée	avec durée
11 mfcc+11Δ+ΔE+ΔΔE	214 (2,4%)	190 (2,2%)
11 mfcc+11Δ+11ΔΔ+ ΔE+ΔΔE	206 (2,4%)	181 (2,1%)

Dans une deuxième expérience, nous avons utilisé les noms épelés du corpus BDSONS. Chaque lettre est modélisée par un HMM2 à 6 états. Un modèle à deux états est utilisé pour le silence. Cette expérience a montré que le modèle de durée permet de réduire significativement le taux d'insertions de lettres (cf tableau 3.). Les taux de subsitution et d'omission restent comparables.

Tableau 3 : taux d'insertion de lettres sur BDSONS

Paramétrisation	sans durée	avec durée
12 mfcc	12%	9%

7. CONCLUSION

L'idée de s'intéresser aux modèles de Markov du second ordre semblait au départ un exercice de style. Les expériences présentées dans cet article montre que ces modèles permettent une meilleure modélisation de la parole. Cette amélioration est obtenue par une prise en compte du fait que le processus de reconnaissance ne peut pas transiter par un état sans y séjourner un nombre non nul de fois, ce qui nous fait dire que ces états ne peuvent pas être transitoires. Comme dans le cas des HMM1, l'ajout des coefficients dynamiques permet une amélioration des performances mais les ΔΔ n'apportent aucune aide comme déjà noté dans le cas de la parole non bruitée. Comme dans les HMM1, il est nécessaire d'éliminer les alignements incorrects que l'algorithme de Viterbi peut construire. Nous avons montré que la classification par mots de tous les alignements corrects permettait de réduire très facilement les taux d'erreurs de 10% ce qui nous fait dire qu'il y a encore beaucoup à gagner en utilisant d'autres sources de connaissance.

8. BIBLIOGRAPHIE

[Carré84] "the french Language database: defining, planing and recording a large database", R. Carré, R. Descout, J.J. Mariani, M. Eskenazi, M Rossi, ICASSP 1984 42.10

[Digalakis91] "A dynamical system approach to continuous speech recognition", V. Digalakis J.R. Rohlicek and M. Ostendorf, ICSSAP 1991, pp 289-292.

[Hanson90] "Robust speaker independent word recognition using static and dynamic and acceleration features: experiments with Lombard and noisy speech" Brian Hanson and Ted Applebaum, ICASSP 1990, pp 857-860.

[Kriouile90] "La reconnaissance automatique de la parole et les modèles markoviens cachés", Abdelaziz Kriouile, thèse de l'université de Nancy I, 1990

[Leonard84] "A database for seaker independent digit recognition", R. G. Leonard, ICASSP 1984, 42.11.1-4

[Levinson86] "continuously varaible duration hidden Markov models for automatic speech recognition" S.E. Levinson, Computer Speech and Language 1986, V1, pp 29-45.

[[Lokbani93] " Segmental post-processing of the N-best solutions in a speech recognition system" M. N. Lokbani, D. Jouvet, J. Monné. Eurospeech 93 , pp 811-814.

[Mari94] "Quelques aspects des modèles de markov en reconnaissance de la parole" ,Jean-François MARI, Dominique FOHR , Jean-Paul HATON. Séminaire " Reconnaissance automatique de la parole", CRIN/ INRIA-Nancy, 10-11 Mars 1994.

[Saerens93] "A continuous-time dynamic formulation of Viterbi algorithm for one-Gaussian-per-state hidden Markov models", Marco Saerens, Speech Communication 1993, V12, pp 321-333.

[Schwartz91] "A Comparison of Several Approximate Algorithms for Finding Multiple (N-BEST) Sentence Hypotheses", Richard Schwartz and Steve Austin, ICASSP 1991, pp 701-704.

[Suaudeau93] "Sound duration modelling and time varianble speaking rate in a speech recognition system", N. Suaudeau, R. André-Obrecht, Eurospeech 1993, pp 307-310.

[Wellekens87] "Explicit time correlation in hidden Markov models for speech recognition" C. Wellekens, ICASSP 1987, pp 384-386.

[Wilpon93] "Connected digit recognition based on improved acoustic resolution", J.G. Wilpon, C-H. Lee and L.R. Rabiner, Computer Speech and Language 1993, V7 pp15-26.

EFFICACITE DES METHODES DIAGNOSTIQUES DE L'INSUFFISANCE VELAIRE SUR SEGMENTS VOCALIQUES.

E. PERRIN, C. BERGER-VACHON, F. PLANTE & I. KAUFFMANN.

Laboratoire Audition et Voix
URA CNRS 1447
Hôpital Edouard-HERRIOT
69437 LYON Cedex 03
E-mail: eperrin@cimacpcu.univ-lyon1.fr

Résumé Cette étude porte sur les modifications spectrales observées dans la phonation de huit voyelles du français entre un groupe de 21 enfants atteints d'insuffisance vélaire et un groupe témoin de 42 enfants. Le matériel phonétique est composé de 3 voyelles isolées (/a/, /i/ et /u/), d'un /a/ tenu, de deux voyelles nasales /ɛ̃/ et /ɔ̃/ dans le contexte du mot dindon et de deux voyelles orales /a/ et /o/ dans le contexte du mot gâteau. Quatre méthodes classiques de traitement du signal de parole ont été utilisées : détection automatique des formants, analyse cepstrale, analyse FFT suivant une échelle linéaire et analyse FFT suivant une échelle bark. A l'aide d'une analyse discriminante, on a montré que les méthodes cepstrales et les méthodes FFT linéaires sont les meilleurs outils pour classer les groupes d'enfants vis à vis de la pathologie. L'utilisation d'une échelle perceptuelle en bark n'apporte pas de meilleurs résultats. En outre, la reconnaissance de la pathologie est améliorée lorsque les sous-populations de garçons et de filles sont considérées. Les meilleurs résultats ont été obtenus pour le /a/ de gâteau pour les filles et pour la voyelle isolée /i/ pour les garçons.

Mots-clés : insuffisance vélaire, voyelles du français, enfant, spectre, cepstre, analyse discriminante.

1. INTRODUCTION

L'étude de la pathologie de la voix et l'aide au handicap sont des directions importantes de la recherche acoustique. Ce travail porte sur la caractérisation acoustique de l'insuffisance vélaire. L'évaluation qualitative de la pathologie peut être une aide précieuse apportée au praticien autant au point de vue du diagnostic décisionnel que du suivi thérapeutique.

La pathologie étudiée ici, l'insuffisance vélaire, est généralement due à un dysfonctionnement ou à un voile du palais trop court. En première approximation, le voile du palais est un "clapet" qui permet soit d'orienter une partie du flux d'air vers les cavités nasales (position ouverte) soit d'isoler ces cavités. Le sujet atteint d'insuffisance vélaire présente une fuite d'air par les cavités nasales, fuite qui entraîne une phonation nasalisée sur les voyelles et des perturbations lors des phénomènes plosifs (ce cas est celui de la rhinolalie ouverte). A l'extrême, le sujet peut compenser cette impossibilité de fermer le conduit nasal avec des coups de glotte et en amenant la langue dans une position exagérément postérieure. L'objet de cette étude est de caractériser cette pathologie à partir des phénomènes quasi-stationnaires de la parole. Ceci est très important dans le cas de l'enfant qui s'accommode mal d'examens agressifs, et les méthodes acoustiques ont alors une place de choix.

L'évaluation acoustique de l'insuffisance vélaire est un sujet difficile et non résolu où le choix du matériel phonétique a son importance [Kingholtz, 1990]. Néanmoins on sait que quelques modifications ont déjà été individualisées [Dickson, 1962; Castelli, 1989] :

- apparition de résonances nasales entre les positions normales des formants,
- création d'antirésonances dans le spectre,
- augmentation de la bande passante des formants.

Divers éléments sont à considérer lors de la reconnaissance de l'insuffisance vélaire : la numérisation et la segmentation des signaux ainsi que le test statistique doivent être adaptés comme dans chacune des études de ce type. Il faut aussi tenir compte de problèmes courant en pathologie vocale, notamment en ce qui concerne l'insuffisance vélaire, qui sont :

- l'influence des voyelles étudiées [Mayo et col., 1993],
- l'influence des méthodes de traitement du signal utilisées [Gramming, 1988; Plante 1993a].

Dans ce travail, on se propose de rechercher l'efficacité de la détection de la pathologie en fonction de la voyelle étudiée et en fonction des paramètres calculés sur le signal vocal. Pour cela, huit voyelles ont été considérées dans le protocole standard utilisé pour l'enfant à l'Hôpital Edouard-Herriot de Lyon : trois voyelles isolées, quatre voyelles en contexte dans un mot (dont deux nasalisées) ainsi qu'une voyelle tenue. Les paramètres considérés sont :

- le fondamental de la voix et les trois premiers formants,
- les coefficients cepstraux issus d'une analyse LPC (LPCC),
- le spectre FFT étudié sur douze bandes d'égale largeur,
- le spectre FFT étudié sur douze bandes de fréquences suivant l'échelle bark.

L'efficacité de la discrimination entre les population normales et pathologiques sera présentée et discutée dans chacun des cas. Ce travail s'intègre dans le cadre plus général des méthodes d'évaluation utilisées en acoustique pour l'étude de phénomènes ciblés.

2. METHODOLOGIE MATHEMATIQUE

2.1 Méthodes de caractérisation

Les méthodes d'évaluation de la pathologie que nous avons étudiées proviennent de l'étude du signal acoustique. Les méthodes d'analyse sont soit issues de la modélisation (LPC), soit des décompositions du spectre en fréquence (FFT) pondérées ou non par des considérations sur la perception du spectre par l'oreille (échelle bark). Dans tous les cas, nous avons privilégié des mises en oeuvre automatiques des algorithmes au moyen du logiciel I.L.S. (Interactive Laboratory System), afin de réduire l'influence du manipulateur sur les résultats et sur les solutions proposées [Vecsys, 1989].

2.1.1 Paramètres généraux

Le signal acoustique a été échantillonné sous 10 kHz à travers un filtre antirepliement de fréquence de coupure 4 kHz. Un préfiltrage passe haut (coupure 80 Hz) a aussi été analogiquement effectué dans le but d'éliminer les basses fréquences parasites. La quantification a été effectuée sur 12 bits (carte d'acquisition DT2825). Les paramètres généraux d'analyse sont présentés dans le tableau 1. Les analyses suivantes ont été effectuées sur les échantillons du signal de parole. Un jeux de coefficients étant obtenu pour chaque fenêtre d'analyse, les résultats ont donc été moyennés sur cinq fenêtres (soit 60 ms) pour chaque voyelle.

Taille de la fenêtre	200 points (20 ms)
Recouvrement	50 %
Préemphase	98%
Fenêtre de pondération	Hamming

Tableau 1. Paramètres généraux de l'analyse.

2.1.2 Fondamental de la voix et analyse LPC.

Nous avons défini un jeu de paramètres en prenant la source glottale (F0) et les trois premiers formants.

Le fondamental de la voix est calculé par le logiciel I.L.S. selon une méthode cepstrale à deux niveaux. Le premier niveau est une détection du pic du fondamental dans le cepstre; le second niveau est une vérification de l'intervalle d'acceptation de ce pic en fonction des analyses précédentes.

La modélisation LPC a été calculée au douzième ordre, ce qui permet de définir cinq formants en dessous de 5 kHz (Fe=10 kHz). Les formants ont été définis à partir de ce modèle et nous n'avons pas discuté les valeurs données par le logiciel I.L.S.

2.1.3 Analyse cepstrale.

Les coefficients cepstraux calculés ici ont été obtenus à partir des coefficients auto-régressifs d'une modélisation LPC du douzième ordre (LPCC). Ce jeu de paramètres comprend donc douze éléments.

2.1.4 Analyse par bandes d'égale largeur sur le spectre FFT (FFT linéaire)

Le spectre calculé à l'aide d'un algorithme de FFT a été divisé en douze bandes de fréquence d'égale largeur. Pour une fréquence d'échantillonnage de 10 kHz et une taille de fenêtre d'analyse de 200 points, le spectre s'étend de 80 Hz à 4 kHz, soit 88 points; chaque bande de fréquence contient donc 8 points sur le spectre, représentant une étendue fréquentielle de 400 Hz.

2.1.5 Analyse par échelle bark sur la FFT (FFT bark)

Pour cette analyse, nous avons divisé le spectre calculé par la FFT en douze bandes de fréquences suivant l'échelle bark. La bande passante utile s'étend de 80 Hz (0.689 bark) à 4 kHz (17.359) bark. Les bornes des bandes sont indiquées dans le tableau 2.

bark	0.689	2.060	3.451	4.841	6.232	7.623	9.014
Hz	80	207.4	342.8	492.1	661.9	859.3	1093

bark	10.405	11.796	13.186	14.577	15.968	17.359
Hz	13.73	1711	2123	2626	3241	4000

Tableau 2. Décomposition du spectre suivant une échelle bark.

2.2 Analyse statistique

La méthode statistique utilisée pour classer les populations témoin et pathologique est l'analyse discriminante. Chaque classement des populations par l'analyse discriminante a été effectué séparément pour chacun des quatre jeux de paramètres calculés, afin de rendre possible les comparaisons des efficacités entre les jeux de paramètres.

Dans l'analyse des résultats, nous avons considéré le pourcentage de bons classements ainsi que le D^2 de Mahanalobis indiquant la séparation des populations. Cette indication permet de comparer la puissance discriminatoire du jeu de paramètres considéré.

3. ACQUISITION DES DONNEES

3.1 Station d'étude

Les analyses menées dans cette études ont été réalisées au moyen d'une station S.A.M. (Speech Assessment Management), pilotée par le logiciel I.L.S.

3.2 Protocole et matériel phonétique

Chacun des enfants de cette étude a été enregistré selon un protocole standard comportant une liste de mot phonétiquement adaptés à l'étude de l'insuffisance vélaire. Les voyelles prises en compte dans cette étude sont indiquées dans le tableau 3.

Voyelle	Contexte
/a/	isolé
/i/	isolé
/u/	isolé
/ɛ̃/	dindon
/õ/	dindon
/a/	gâteau
/o/	gâteau
/a/	tenu début
/a/	tenu milieu
/a/	tenu fin

Tableau 3. Voyelles étudiées

3.3 Populations

21 enfants porteurs d'insuffisance vélaire ont été inclus dans cette étude; il y a 10 filles et 11 garçons. Le degré de pathologie a été évalué par le médecin phoniatre du service selon l'échelle Borel-Maisonny [1975] (Tableau 4a). Cette échelle classe les phonations associées à l'insuffisance vélaire en quatre stades :

Stade I : phonation normale,
Stade IIa : légère nasalité,
Stade IIm : forte nasalité,
Stade III : phénomènes compensatoires.

	Stade IIa	Stade IIm	Stade III
Garçons	5	4	2
Filles	3	6	1
Total	**8**	**10**	**3**

Tableau 4a. Répartition des enfants porteurs d'insuffisance vélaire en fonction du degré de pathologie.

Deux sujets témoins ayant le même sexe et le même âge ont été associés à chacun des enfants pathologiques à partir de la banque de données du laboratoire (Tableau 4b). La population totale de cette étude est de 63 enfants.

Age	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Garçons	4 *(8)*	1 *(2)*	1 *(2)*		3 *(6)*	1 *(2)*	**10** ***(20)***
Filles	3 *(6)*	1 *(2)*	4 *(8)*	1 *(2)*	1 *(2)*	1 *(2)*	**11** ***(22)***

Tableau 4b. Sexe et âge (en années) des enfants de l'étude porteurs d'insuffisance vélaire. Entre parenthèses, on indique le nombre d'enfants témoins correspondants.

4. RESULTATS GLOBAUX ET DISCUSSION

4.1 Résultats par population

Les résultats obtenus sur la population totale (63 sujets) sont présentés sur la figure 1. Les meilleurs pourcentages de discrimination ont été obtenus avec :

- la voyelle /a/ et le cepstre (94 %, $D^2=5.2$),
- la fin su /a/ tenu et la méthode FFT linéaire (92 %, $D^2=9.6$),
- le début du /a/ tenu et le cepstre (92 %, $D^2=4.9$).

Les meilleurs résultats sur la population totale se situent au delà de 90 %, ce qui correspond à un mauvais classement de moins de 6 sujets sur 63.

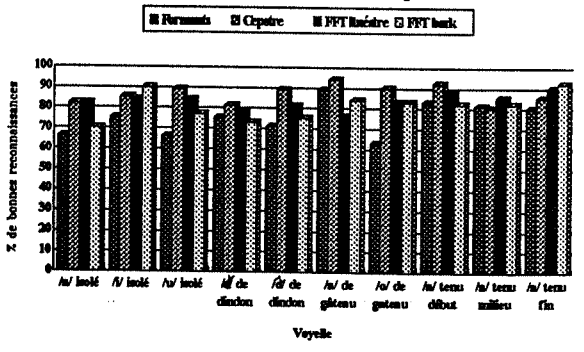


Figure 1. Pourcentages de bonnes reconnaissances obtenus pour la population totale.

Un autre classement par analyse discriminante a été effectué en considérant les populations de filles

(30 sujets, figure 2) et de garçons (33 sujets, figure 3).

Population des filles :

Sept scores égaux à 97% ont été obtenus, ce qui correspond au mauvais classement d'un seul sujet (Figure 2). Ces scores sont obtenus avec :

- la voyelle /ɑ̃/ et le cepstre ($D^2=11.2$),
- le début du /a/ tenu et le cepstre ($D^2=12.2$),
- la voyelle /u/ tenu et le cepstre ($D^2=9.8$),
- le fin du /a/ tenu et la FFT linéaire ($D^2=9.6$),
- la voyelle /u/ et la FFT linéaire ($D^2=7.4$),
- le milieu du /a/ tenu et la FFT linéaire ($D^2=9.6$),
- la voyelle /i/ et la FFT linéaire ($D^2=6.8$).

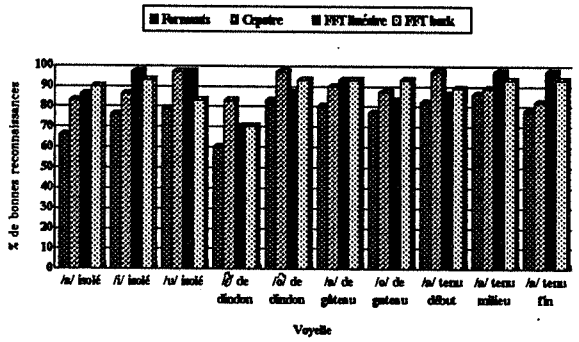


Figure 2. Pourcentages de bonnes reconnaissances obtenus chez les filles (30 sujets).

Population des garçons :

Trois analyses ont montré 100% de bons classements (Figure 3) avec :

- la voyelle /a/ isolée et le cepstre,
- la voyelle /i/ isolée et le cepstre,
- la voyelle /i/ isolée et le cepstre.

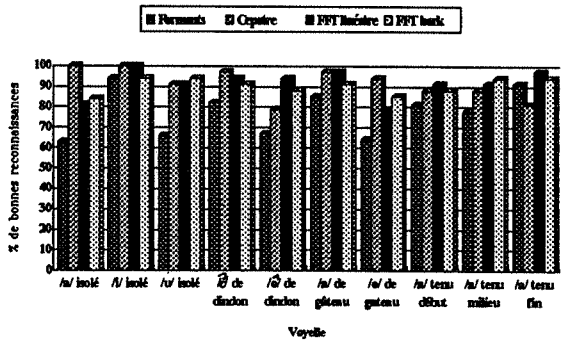


Figure 3. Pourcentages de bonnes reconnaissances obtenus chez les garçons (33 sujets).

Ces résultats considérés par sexe conduisent à de meilleurs pourcentages de reconnaissance ainsi qu'à une amélioration du pouvoir de discrimination (augmentation du D^2 de Mahanalobis) par rapport au classement effectué sur la population totale. Ces résultats sont bons puisqu'ils atteignent 100% chez les garçons et 97% chez les filles ce qui correspond au mauvais classement d'un seul sujet.

4.2 Résultats par méthode

Le tableau 5 montre une synthèse des résultats obtenus par les quatre méthodes. Les méthodes FFT et la méthode cepstrale donnent globalement des résultats équivalents, clairement supérieurs aux formants.

	Formants	Cepstre	FFT linéaire	FFT bark
Filles	78%	88%	89%	89%
Garçons	77%	91%	91%	90%
Total	75%	87%	83%	81%
Moyenne	77.6%	86.6%	87.6%	86.6%

Tableau 5. Pourcentages globaux de bonne reconnaissance par méthode et par population

4.3 Résultats par voyelle

L'efficacité des voyelles pour la séparation des populations est indiquée sur la figure 4. Les meilleurs résultats ont été obtenus avec le /a/ de gâteau, le début et la fin du /a/ tenu.

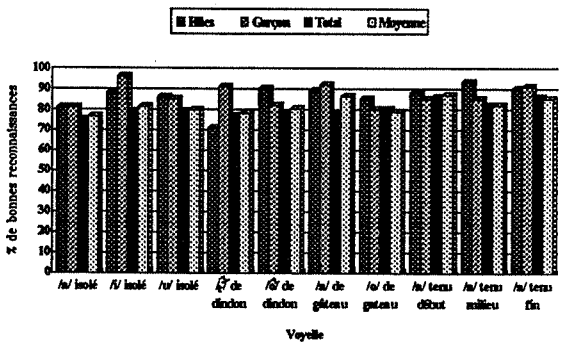


Figure 4. Pourcentages de séparation obtenus avec les voyelles.

La significativité de ces constatation méritera d'être envisagée dans de futures études.

5. CONCLUSION

Les conclusions sur cette étude peuvent être envisagées selon deux angles d'analyse.

Les résultats sont encourageants puisqu'ils permettent un classement avec moins de 10% d'erreur pour une population totale de 63 sujets. Si l'on effectue le classement en considérant à part les

garçons et les filles, les pourcentages de reconnaissance sont améliorés jusqu'à 100% chez les garçons et 97% chez les filles.

En ce qui concerne l'efficacité des quatre méthodes, l'analyse FFT sous ses deux formes et l'analyse cepstrale, se sont montrées les plus performantes que l'utilisation des formants déterminés automatiquement et ce dans presque tous les cas. L'introduction de l'échelle bark n'a pas apporté une nette amélioration des résultats.

Le choix des voyelles n'est pas non plus indifférent; dans ce travail, le /a/ sous toutes ses différentes formes, a conduit aux meilleurs résultats.

REMERCIEMENTS

Les auteurs remercient les organismes qui les soutiennent dans leur travail, les Hospices Civils de Lyon, la Région Rhône-Alpes, la société API de Lyon, le CNRS et l'université Claude Bernard, ainsi que MM S. Gallégo et S. Garnier pour le travail qu'ils ont fourni lors du dépouillement des données.

Ils remercient aussi les professeurs L. COLLET et A. MORGON (Service d'O.R.L. de l'Hôpital Edouard-Herriot) pour les moyens qu'ils ont mis à leur disposition.

BIBLIOGRAPHIE

S.BOREL-MAISONNY "L'insuffisance vélaire; point de vue orthophonique", Rev. Orthophonie 13, [61-68], 1975.

E. CASTELLI "Caractérisation acoustique des voyelles du français", Thèse de Sciences, INP Grenoble, 1989.

J. CHERPILLOD & M.L. DUTOIT-MARCO " Bilan fonctionnel de l'insuffisance vélaire", Médecine et hygiène, 40, [3739-3740], 1982.

R.M. DALSTON, D.W. WARREN & E.D. DALSTON " Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairement", Cleft Palate J., 28, [184-189], 1991.

D.R. DICKSON "An acoustic study of nasality", J. Speech Hear. Research, 5, [103-111], 1962.

E. DILLENSCHNEIDER, T. ZALESKI & G.F. GREINER "Etude sonagraphique de la nasalité dans les cas d'insuffisance palato-vélaire", J. F. O.R.L., 22, [201-202], 1973.

J.L. FLANAGAN "Speech analysis, synthesis and perception", Spinger Verlag, 1972.

P. GRAMINA & J. SUNBERG "Spectrum factors relevant to phonetogram measurements", J. Acoust. Soc. Am., 83, [2352-2360], 1988.

Y. HORII "Jitter and Schimmer differences among sustained vowels", J. Speech Hear. Research, 25, [12-14], 1991.

F. KLINGHOLTZ "Acoustic recognition of voice disorders: a comparative study of running speech versus sustained vowels", J. Acoust. Soc. Am., 87, [2218-2224], 1990.

R. MAYO, R.M. DALSTON & D.W. WARREN "Perceptual assessment of resonance distortion in unoperated cleft palate", Cleft Palate J., 30, [397-400], 1993.

F. PLANTE "Détection acoustique des pathologies phonatoires chez l'enfant", Thèse de Sciences, INP Grenoble, 1993a.

F. PLANTE, C. BERGER-VACHON & I. KAUFFMANN "Acoustic determination of velar impairement in children", Folia Phoniatrica, 43, [112-119], 1993b.

G.S. SEBESTYEN "Decision making processes in pattern recognition", Mc Millan, 1962.

VECSYS (STE) "Le système Unice", Le Chêne Rond, 91570, Bièvres, 1989.

D.W. WARREN "The determination of velopharyngeal incompetence by aerodynamic and acoustical techniques", Clin. In Plast. Surg., 2, [299-304], 1975.

E. ZWICKER & H. FASTL "Psychoacoustics: facts and models", Springer Verlag, 1990.

Un environnement logiciel pour un atelier phonétique

Yves Laprie
CRIN – CNRS / INRIA Lorraine
Campus scientifique, B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX

Laurent Mercier
Institut de Phonétique de Nancy
Université de Nancy II, B.P. 3397
54015 Nancy CEDEX

France (email: laprie@loria.fr mercier@loria.fr)

Résumé

La reconnaissance automatique de la parole continue multilocuteurs nécessite l'utilisation d'un volume important de connaissances acoustiques et phonétiques. Nous avons développé Snorri, un environnement logiciel destiné à aider le chercheur en parole à construire un système de décodage acoustico-phonétique (DAP) utilisant explicitement ces connaissances. Snorri offre, outre les outils classiques d'investigation du signal de parole, un ensemble d'outils spécifiquement destinés à une approche analytique du DAP : des extracteurs automatiques d'événements acoustiques (algorithme de suivi automatique des formants, analyseur de la barre d'explosion...) et un environnement permettant de représenter, d'éditer et d'échanger les indices acoustiques avec un système de DAP. Le développement de Snorri a accompagné celui des systèmes de DAP à base de connaisances conçus au Crin et il n'offre donc pas forcément un environnement logiciel adapté aux autres chercheurs en parole. Nous développons donc actuellement un atelier phonétique plus modulaire qui permet à l'utilisateur d'organiser les outils existants en fonction de son travail de recherche.

1 Introduction

Un certain nombre de systèmes de dictée automatique sont commercialisés, ou sur le point de l'être, pour des machines destinées au grand public. Généralement, afin d'offrir un niveau de fiabilité suffisante pour être utilisés par le plus grand nombre, ces systèmes reposent sur l'utilisation de modèles de Markov cachés. Plus que les gains de puissance offerts par les nouvelles générations de puces informatiques, c'est sans doute la prise en compte de connaissances de parole qui explique le succès de ces modèles. En effet, les mécanismes de production de la parole sont si complexes que la voie la moins difficile pour introduire des connaissances dans un système de reconnaissance automatique consiste à modifier le comportement théorique des modèles de Markov ou de soumettre aux algorithmes d'apprentissage de ces modéles un corpus suffisamment vaste pour qu'il couvre le mieux possible les sources de variabilité de la parole.

La poursuite de la voie ouverte par les systèmes à base de connaissances (notamment Aphodex au Crin [Carbonell 84]) dans lesquels les connaissances sont introduites explicitement se heurtent à deux obstacles majeurs. D'une part, il faut expliciter les connaissances acoustiques et phonétiques, puis vérifier l'intérêt que représente leur prise en compte pour la reconnaissance automatique. D'autre part, il faut développer les extracteurs automatiques d'indices acoustiques manipulés par ces connaissances, et les valider sur des corpus de parole suffisamment vastes.

Pour résoudre ces deux problèmes, nous estimons qu'il faut offrir aux chercheurs en parole un environnement logiciel spécifiquement destiné au développement d'une approche à base de connaissances du décodage acoustico-phonétique. À cet effet, nous avons commencé à développer, en 1987, le logiciel Snorri [Fohr 89] que nous utilisons intensivement, en particulier dans le cadre d'un projet portant sur la reconnaissance de phrases construites d'occlusives et de voyelles [Bonneau 92].

Dans cet article nous montrerons d'abord comment nous nous servons de Snorri pour mettre au point les connaissances du système Daphné [Bonneau 92]. Comme le développement de Snorri a accompagné étroitement le développement d'Aphodex et maintenant de Daphné, il n'est pas forcément parfaitement adapté à d'autres recherches sur la parole. Nos efforts s'orientent maintenant vers la mise au point d'un atelier phonétique dans lequel chaque outil correspond à une tâche bien précise (édition du signal de parole, calcul de spectrogramme...) et est facilement interfaçable avec les autres outils disponibles. La description de cette nouvelle approche et des outils déjà réalisés fera l'objet de la seconde partie de l'article.

2 Utilisation de Snorri

Puisque Snorri doit nous aider à développer un système de DAP utilisant des connaissances phonétiques, nous attachons une très grande importance aux détecteurs automatiques des événements acoustiques qui servent en reconnaissance. Nous avons ainsi doté Snorri des trois outils suivants dont nous donnerons une brève description :

- un algorithme automatique de suivi de formants utilisé pour identifier les voyelles et déterminer les transitions formantiques,
- un détecteur et analyseur du bruit d'explosion des occlusives,
- un détecteur de fréquence fondamentale.

Pour mettre au point ces détecteurs nous avons pleinement exploité les possibilités de la version de base de Snorri :

- l'exploration automatique d'un corpus étiqueté pour étudier les contextes phonétiques susceptibles de poser des problèmes, comme par exemple les voyelles d'arrière lors d'un suivi de formants,
- les outils de visualisation de coupes spectrales pour analyser les causes d'erreur du suivi ou les résultats de l'analyseur de barres d'explosion.

Le fait de disposer de ces outils permet d'analyser et d'expliquer les causes d'échec des extracteurs d'indices puis de corriger les algorithmes pour qu'ils prennent en compte plus finement les spécificités du signal de parole.

2.1 Les outils de base de Snorri

La version de base de Snorri tourne sur les postes de travail Unix sous X11 de base et offre les outils de base pour l'étude de la parole :

- éditeur de signal (enregistrer, écouter, couper, coller...) qui permet de manipuler la plupart des signaux de parole,
- calcul de spectrogrammes (à bande large ou étroite, par LPC ou lissés cepstralement).
- calcul de coupes spectrales diverses, affichage de pics spectraux sur le spectrogramme.
- étiquetage phonétique de la parole.
- copies d'écran...

Une version tournant avec OSF/Motif est en cours de développement.

2.2 Suivi de formants

Le raisonnement adopté dans Daphné [Bonneau 92] est de nature abductive puisqu'il consiste à expliquer les observations acoustiques en fonction de la modélisation des sons en termes d'indices acoustiques. Le principe de l'algorithme de suivi est globalement le même, puisqu'il s'agit de détecter les lignes de pics du spectrogramme (obtenu par LPC ou lissage cepstral) puis de les interpréter en termes de formants de la cavité orale. L'interprétation est guidée (et contrainte) par les valeurs formantiques des voyelles car il n'existe que peu de données formantiques pour les consonnes. Elle est donc d'autant plus sure que les sons analysés sont des voyelles orales.

Par ailleurs, dans le cas de l'identification des occlusives, il est important que le suivi permette d'évaluer correctement les transitions formantiques qui suivent le bruit d'explosion. Pour cela l'interprétaion des lignes de pics est suivie d'une méthode de régularisation [Berger 92] qui déforme itérativement les lignes de pics pour réaliser le meilleur compromis entre la régularité des trajectoires formantiques (nécessaire pour calculer précisément les pentes) et la proximité des trajectoires aux lignes de pics (Fig. 1).

Finalement, comme le système de DAP utilisant le suivi vérifie la cohérence des différents indices (formants et caractéristiques de la barre d'explosion) l'algorithme de suivi peut retourner plusieurs suivis : ceux ayant donné lieu aux meilleurs scores d'évaluation. De cette façon le système de DAP peut utiliser le caractère redondant des indices acoustiques en les vérifiant mutuellement et ne retenir que les combinaisons d'indices cohérentes.

2.3 Détecteur de la fréquence fondamentale

Un algorithme de détection de F0 a deux utilisations possibles en reconnaissance automatique: la segmentation de la parole et la détermination d'indices prosodiques. Dans le cadre du projet Daphné, il était nécessaire de disposer très précisément du début du voisement qui suit le bruit d'explosion. En effet, les caractéristiques de la barre d'explosion sont obtenues après que le bruit d'explosion a été segmenté en barre d'explosion et bruit de friction comme nous le montrerons dans le paragraphe suivant.

Nous avons donc développé un algorithme utilisant le peigne fréquentiel proposé par P. Martin [Martin 81] et un algorithme de correction issu de l'algorithme proposé par H. Ney [Ney 83]. En ce qui concerne le peigne fréquentiel, nos efforts ont porté sur la précision fréquentielle des candidats possibles pouvant représenter F0. Pour l'algorithme de correction, nous avons accordé beaucoup d'importance à la détermination du début de voisement. Les résultats

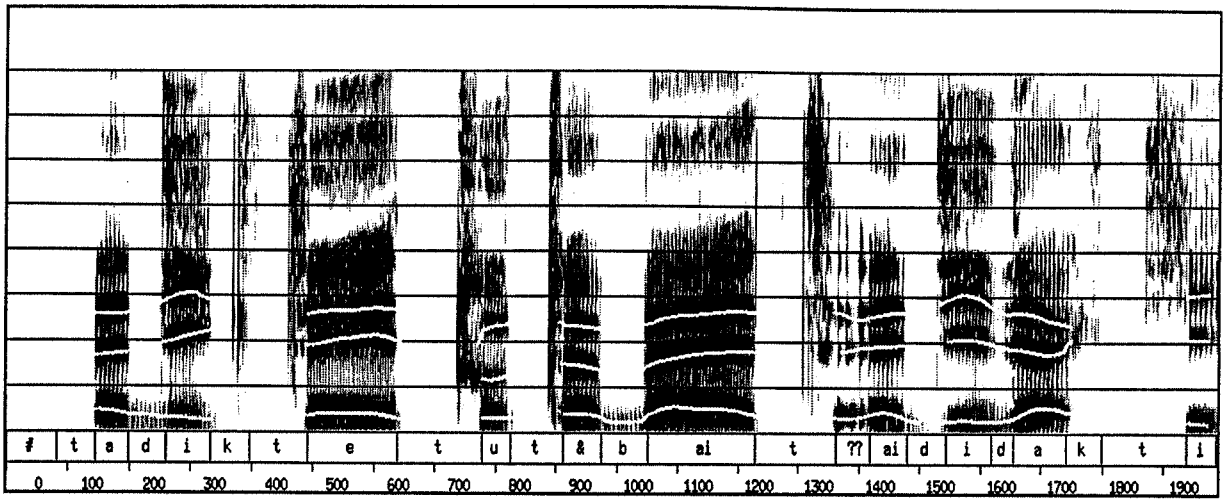


FIG. 1 - *Suivi automatique de formants sur « Ta dictée toute bête est didactique »*

sont tout à fait satisfaisants pour les différentes parties du corpus BDSONS. Nous travaillons actuellement à optimiser la détection du début de voisement sur de la parole enregistrée dans de moins bonnes conditions (radio, bureau calme) et à la détection des « accidents » (doublements de fréquence réels, voix craquée) de la courbe de F0.

2.4 Segmentation et analyse du bruit d'explosion

L'identification des occlusives dans Daphné fait appel aux transitions formantiques et aux caractétistiques spectrales de la barre d'explosion qui constituent des informations essentielles dans la plupart des contextes vocaliques. Les caractéristiques spectrales sont obtenues en deux étapes :

- segmentation du bruit d'explosion en deux parties : barre d'explosion et bruit de friction,
- analyse du spectre de la barre d'explosion.

Le bruit d'explosion s'étend depuis l'apparition de l'énergie jusqu'au début du voisement. La segmentation s'effectue de la manière suivante. On fait l'hypothèse que le bruit d'explosion se décompose en deux parties, chacune représentée par son spectre moyen, la frontière entre les deux correspond à l'instant maximisant, pour les deux parties, la somme des corrélations entre le spectre moyen et les spectres à large bande de la partie considérée ; éventuellement, c'est par exemple le cas pour /p/ où le bruit d'explosion est très faible, l'une des deux parties peut disparaître.

Les caractéristiques spectrales sont ensuite obtenues à partir du spectre moyen de la barre d'explosion en recherchant les pics d'énergie du spectre moyen et en évaluant sa compacité.

2.5 Mise au point d'indices acoustiques

L'identification des sons dans un système de DAP analytique repose sur l'extraction automatique d'un certain nombre d'indices acoustiques qui doivent satisfaire deux conditions :

- l'extraction de ces indices doit être aussi fiable que possible et il faut donc pouvoir tester les extracteurs sur des corpus de parole étiquetés.
- le pouvoir de discrimination de ces indices doit être le meilleur possible. Il faut donc pouvoir récupérer ces indices dans un système de DAP afin de les valider.

Snorri offre la possibilité d'extraire automatiquement des séquences de sons d'un corpus étiqueté, ce qui permet de tester un indice acoustique sur le type même de sons auxquels il est destiné.

L'interfaçage avec le système de DAP est réalisé par l'intermédiaire de la structure de triplet qui représente un son en contexte. Le triplet est organisé autour des deux frontières du son central auxquelles sont attachées les différents indices acoustiques (formants, caractéristiques de la barre d'explosion...). La Fig.2 représente un triplet occlusif. Pour qu'il soit possible d'interagir avec le niveau d'extraction des indices comme avec le niveau d'identification des sons, il est possible d'éditer ces triplets, pour corriger une erreur de détection acoustique, ou, au contraire, pour observer comment le système de DAP se comporte en présence de données acoustiques erronnées.

L'édition des triplets est faite dans Snorri en « dessinant » directement avec la souris sur le spectrogramme les différents événements acoustiques caractérisant le son étudié. Bien sûr, les extracteurs automatiques peuvent être utilisés pour gagner du

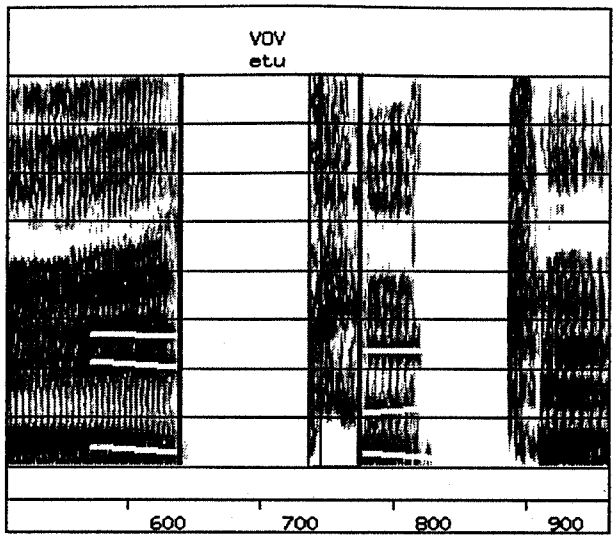


FIG. 2 - *Triplet occlusif muni de ses différents indices acoustiques*

temps. Les triplets sont sauvegardés sous la forme d'objets Lisp avant d'être utilisés par le système de reconnaissance automatique développé avec X-TRA [Charpillet 88].

3 Vers un atelier de phonétique

Comme nous venons de le montrer, Snorri est fortement lié au développement d'un système de DAP. Par conséquent, il n'est pas forcément adapté aux besoins d'autres chercheurs en parole. Au lieu de construire directement un système d'étude de la parole comme Snorri, il est possible de spécifier les types d'objets manipulables et de définir les opérateurs associés à ces objets. Assia [Gong 85], ou ISP [Kopec 84] relèvent de cette approche puisqu'ils reposent sur la définition de types abstraits de base, parmi lesquels le signal de parole. Le chercheur peut enrichir son système en définissant de nouveaux types et les opérateurs associés. Cette approche intéressante a cependant le défaut de demander un certain effort au chercheur pour développer un environnement de travail répondant à ses besoins.

Nous avons donc adopté un point de vue intermédiaire entre un système prêt à l'emploi, comme Snorri, et un système à base de types abstraits en développant un ensemble d'outils de traitement de la parole qui communiquent entre eux, échangeant ou partageant des données typées qui sont disponibles d'entrée. L'objectif est de montrer qu'il est plus profitable de construitre un atelier phonétique constitué de plusieurs applications indépendantes mais communiquant entre elles plutôt que d'avoir une seule vaste application peu maîtrisable. Nous expliciterons tout d'abord les concepts de base de l'atelier phonétique puis nous expliquerons sommairement les applications qui forment le noyau de cet atelier.

3.1 Philosophie de l'atelier phonétique

Le principe de base de notre atelier phonétique est la communication entre applications. Il est important de mentionner que ces applications ont été écrites en langage C sous X window en utilisant les fonctions de la librairie « Intrinsics » et le jeu de « widgets » OSF/Motif. Ces applications sont dotées d'un même processus de transfert d'informations. Tout l'atelier phonétique est conçu sur ce modèle de communication. Techniquement, les objets peuvent être transférés d'une application vers une autre, soit sous la forme d'un fichier, soit par l'intermédiaire d'un mécanisme particulier de « couper-coller ». L'atelier phonétique permet donc de faire transiter un ensemble d'informations homogènes, considérées comme un objet, à travers différentes applications. Cet atelier phonétique est évolutif. Il permet de créer de nouveaux objets à partir d'objets déjà définis.

Les opérations de chaque application sont spécifiques à l'objet qu'on manipule et doivent permettre à l'utilisateur d'avoir un certain regard sur celui-ci. Clairement, l'utilisateur peut agir sur les objets par l'intermédiaire des applications. Dans le cas d'extraction d'événements acoustiques, il peut par exemple intervenir manuellement sur un objet pour corriger les erreurs produites par une procédure automatique.

3.2 Description de l'atelier phonétique

La Fig.3 présente les objets d'entrée et les objets de sortie des trois applications de base qui forment notre atelier phonétique. Elle montre qu'un objet de sortie d'une application est l'objet d'entrée d'une autre. L'objet est transformé en un nouvel objet à chaque fois qu'il passe à travers une application. Enfin, elle indique de quelle manière il est possible d'ajouter une nouvelle application et un nouvel objet à l'atelier. Nous ne discuterons pas sur la structure des objets, nous décrirons seulement et rapidement les spécifités des applications Msignal, Mspectro et Mfo.

Msignal est l'éditeur de signal de parole. Celui-ci permet d'afficher et de faire défiler dans une fenêtre le signal de parole tout en contrôlant l'échelle temporelle et l'échelle d'amplitude.

Mspectro permet d'afficher et de faire défiler un spectrogramme dans une fenêtre. Il peut calculer un spectrogramme par FFT ou bien par LPC. Il permet enfin de segmenter et d'étiqueter manuellement la parole, mais aussi, de lancer à tout moment un certain nombre d'analyses spectrales. Mspectro transforme par exemple l'objet *Signal* en un objet *Spectro* qui peut être utilisé pour suivre automatiquement

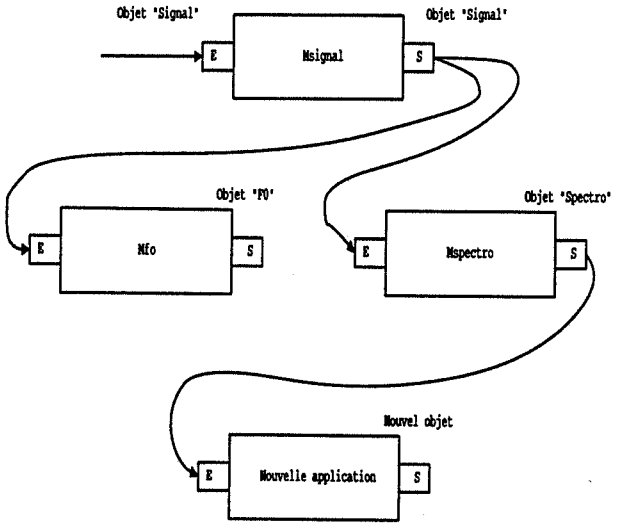


FIG. 3 - *Description de l'atelier phonétique*

des formants. Il suffira de créer l'application capable de suivre automatiquement les formants à partir de l'objet *Spectro*.

Mfo permet d'afficher et de faire défiler la courbe de F0 dans une première fenêtre et un spectrogramme à bande étroite dans une deuxième fenêtre. La fréquence fondamentale peut être calculée selon différentes méthodes. La présence d'un spectrogramme à bande étroite permet de se repérer dans le signal temporel et surtout de vérifier ou de corriger les éventuelles erreurs de calcul de F0.

En utilisant simultanément Mspectro et Msignal, il est possible à l'aide de la souris de sélectionner une portion de signal de parole dans la fenêtre de Msignal puis de la coller dans la fenêtre de Mspectro. Alors le spectrogramme correspondant à la portion de signal de parole sélectionnée s'affichera automatiquement dans la fenêtre de Mspectro. Enfin Mspectro permet de changer les caractéristiques du spectrogramme: choix de la taille de la FFT, ou pour une LPC, le nombre de coefficients de prédiction linéaire.

4 Conclusion

Les deux approches que nous venons de décrire sont destinées à offrir au chercheur en parole un environnement logiciel pour examiner facilement le signal de parole et ainsi l'aider à enrichir les connaissances acoustiques et phonétiques utilisables dans un système de DAP. Pour atteindre cet objectif, proche de celui de Spire [Zue 86], nous pensons qu'il est essentiel de concevoir simultanément les algorithmes pour extraire du signal les événements acoustiques, et les mécanismes de raisonnement qui les interprèteront phonétiquement en termes de sons. En effet, les deux étant utilisés conjointement dans un système de reconnaissance, il faut les valider ensemble, pour vérifier que l'intéraction des deux niveaux ne provoque pas d'erreur de reconnaissance.

Outre la poursuite du développement de Snorri et de l'atelier de phonétique, nous pensons qu'il existe deux directions pour orienter notre travail à venir :

- l'ajustement des paramètres d'extraction des événements acoustiques influençant le pouvoir de discrimination des indices acoustiques. En effet, le chercheur en parole, guidé par ses connaissances des phénomènes de production ou de perception de la parole, choisit des indices acoustiques qui permettent de discriminer les sons entre eux, donc de les identifier. En revanche, il a plus de difficultés à choisir les valeurs des différents paramètres numériques (de fréquence, d'intensité...) qui peuvent modifier sensiblement l'efficacité d'un indice en reconnaissance. Il faut donc lui fournir des outils automatiques lui permettant de réaliser ces règlages au mieux pour un corpus d'entraînement [Zue 89].

- l'utilisation de logiciels dans une équipe de chercheurs en parole conduit inévitablement à l'apparition d'un certain nombre de versions issues d'un même logiciel de base. Il faut donc intégrer des outils de génie logiciel pour gérer les versions concurrentes, ce qui qui permet de faire évoluer le logiciel de manière plus souple.

Remerciements

Nous remercions François Lonchamp pour les conseils qu'il nous a prodigués tant durant le développement de Snorri que durant celui de l'atelier de phonétique.

Nous remercions également Nathalie Flammang qui travaille actuellement au portage de Snorri sous OSF/Motif.

Références

[Berger 92] M. O. Berger et Y. Laprie. Global Active Method for Automatic Formant Tracking Guided by Local Processing. *Proceedings 11th IAPR International Conference on Pattern Recognition*, pages 598–601, The Hague (The Netherlands), September 1992. IEEE Computer Society Press.

[Bonneau 92] A. Bonneau, S. Coste, L. Djezzar et Y. Laprie. Two Level Acoustic Cues for Consistent Stop Identification. *Proceedings International Conference on Spoken Language Processing*, volume 1, pages 511–514,

Banff (Alberta, Canada), October 1992.

[Carbonell 84] N. Carbonell, D. Fohr, J. P. Haton, F. Lonchamp et J. M. Pierrel. An expert system for the automatic reading of french spectrograms. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing 1984*, San Diego, 1984.

[Charpillet 88] F. Charpillet, P. Théret et J.P. Haton. X-TRA : un moteur d'inférence comportant deux modes de compilation des règles TREAT et RETE et un système de maintien de la vérité de type ATMS. *Actes des journées internationales des systèmes experts*, Avignon, Mai 1988.

[Fohr 89] D. Fohr et Y. Laprie. Snorri: an interactive tool for speech analysis. *Proceedings of European Conference on Speech Technology*, Paris, France, September, 1989.

[Gong 85] Y. Gong et J. P. Haton. Manipulation et analyse de signaux sous Unix : l'éditeur Assia. *Une Achitecture Nouvelle et ses Applications, Actes des Journées Sm90*, Paris, 1985. Agence de l'Informatique.

[Kopec 84] G. E. Kopec. The integrated signal processing system ISP. *IEEE Trans. on Acoust., Speech, Signal Processing*, ASSP-32:842–851, August 1984.

[Martin 81] Ph. Martin. Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne. *Actes des 12èmes Journées d'Etudes sur la Parole*, pages 223–232, 1981.

[Ney 83] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173, March 1983.

[Zue 86] V.W Zue, D.S. Cyphers, R.H. Kassel, D.H. Kaufman, H.C. Leung, M. Randolph, S. Seneff, J.E. Unverferth et T. Wilson. The development of the MIT lisp-machine based speech research workstation. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986.

[Zue 89] V.W Zue, J. Glass, M. Phillips et S. Seneff. The MIT SUMMIT speech recognition system : a progress report. *Proc. of the 2nd DARPA Speech and Language Workshop*, 1989.

L'ÉCART ENTRE F1 ET F2 COMME MESURE DE L'EFFET DES CONSONNES LINGUALES SUR LES CIBLES ACOUSTIQUES DES VOYELLES DE L'ARABE STANDARD MODERNE

ZNAGUI Imad & BOUCHNAFI Ahmed

LABORATOIRE DE PHONÉTIQUE DE PARIS III
(CNRS URA 1027 ET UNIVERSITE DE LA SORBONNE NOUVELLE)

Résumé

L'objet de cette étude est l'utilisation d'une approche acoustico-phonétique des voyelles de l'Arabe Standard Moderne (ASM), prononcées par des locuteurs maghrébins. On propose d'évaluer le degré de coarticulation (gauche-droite) des voyelles brèves inaccentuées /a i u/ et longues accentuées /a: i: u:/ précédées des consonnes linguales de différents lieux d'articulation (interdentales, alvéolaires, postalvéolaires, palatales, postpalatales, vélaires-uvulaires, pharyngales et pharyngalisées). Les résultats de l'analyse spectrographique mettent en évidence les conséquences du mouvement d'avancement et de retraction du corps de la langue sur les cibles acoustiques des voyelles en structure CV. Ces mêmes résultats montrent que le degré de la coarticulation linguale varie en fonction de l'aperture fermée ou ouverte de la voyelle. Il en résulte que la voyelle ouverte /a/ est la plus vulnérable à la coarticulation linguale par rapport aux deux voyelles fermées /i/ et /u/. Toutefois, on estime que le critère de distance entre les deux premiers formants (F1 et F2) permet d'élaborer une classification acoustico-phonétique binaire du système vocalique de l'ASM.

Mots-clés: mouvement du corps de la langue, vocoïdes postériorisés, vocoïdes antériorisés, cibles acoustiques, trajectoires formantiques.

INTRODUCTION

Les sons de la parole sont définis par leur lieu d'articulation dans le conduit vocal, de la glotte aux lèvres . En changeant la position d'une articulation à l'autre, on passe d'une région qui peut être sensible ou non aux mouvement articulatoires et/ou aux variations acoustiques. Les nomogrammes de Fant (1960) et ceux de Stevens (1989) montrent clairement la relation entre les deux paramètres articulatoire et acoustique.
Par ailleurs il ressort des études sur la coarticulation (MacNeilage et Declerk (1969), Recasens (1991), Schouten et Pols. (1979) que les mouvements des articulateurs comme le corps de la langue, entre-autres, ne s'arrêtent pas instantanément d'une configuration à une autre et dépassent le segment phonématique. Nous sommes conscients que le degré de coarticulation linguale dépend de la nature de la voyelle, de sa position accentuée et inaccentuée, de sa durée, de la nature de la consonne linguale adjacente.

A ce titre, le système vocalique de l'ASM relève d'un domaine d'investigation intéressant sur le plan de la coarticulation, du fait du nombre élevé de l'entourage des consonnes réalisées avec la langue : 22 consonnes linguales /ð θ t d n l r j s z ʃ ʒ k χ q ʁ ʕ ħ tˤ sˤ ðˤ dˤ/de différents lieux d'articulation (interdentales, alvéolaires, postalvéolaires, palatales, postpalatales, vélaires-uvulaires, pharyngales et pharyngalisées). Il nous a paru, dès lors, intéressant d'examiner le degré des effets de la coarticulation linguale gauche-droite (en structure CV) sur les cibles des voyelles adjacentes brèves inaccentuées /a, i, u / et longues accentuées /a:, i:, u:/. Pour cela, nous avons entrepris une approche acoustico-phonétique où l'écart entre F1 et F2 est pris comme mesure des effets des consonnes linguales de l'ASM sur le timbre des voyelles précitées.

CADRE THÉORIQUE

L'importance de l'écart entre deux formants a été abordée par Chistovitch et Lublinskaya (1979) dans l'analyse des voyelles synthétiques. Ces chercheurs ont constaté que le changement de timbre s'effectue au moment où les deux formants sont plus rapprochés. De même, Ce rapprochement est préféré perceptivement (critère de prégnance). Les voyelles peuvent donc être distinguées entre elles par le phénomène de distribution d'énergie produite par des "focal points". On l'explique par des points de convergence des formants qui servent à la distinctivité auditive des timbres vocaliques comme dans les cas suivants:
- convergence de F1 et F2 de la voyelle /u/ avec les lèvres fermées.
- convergence de F1 et F2 de la voyelle /a/ avec les lèvres ouvertes.
- convergence de F2 et F3 de la voyelle /y/ avec les lèvres arrondies.

- convergence de F3 et F4 de la voyelle /i/ avec les lèvres fermées.

Le phénomène de convergence traduit une instabilité et des discontinuités dans le signal de parole et provoque un échange d'affiliation entre les cavités et les deux formants acoustiques.

Nous pensons que le concept de distance "critique" entre les deux premiers formants permet d'évaluer l'effet des consonnes linguales des différents lieux d'articulation sur le timbre des voyelles de l'ASM. Les travaux de Öhman (1967), Perkell (1969), Lindblom et Sundberg (1971) mettent en évidence le rôle de la position et la forme de la langue dans la coarticulation entre consonne-voyelle et voyelle-consonne.

Dans la même optique, Merlmestein (1973) propose pour la synthèse par règles un modèle dynamique du système articulatoire et des règles de coarticulation :"vowel-consonant-vowel sequences and articulator-position to vocal-tract coarticulation in an agreement with acoustic measurements"(Merlmestein (1973), p.1070).

Boé et al. (1992) classent les paramètres articulatoires en fonction de leur importance dans la production des voyelles. Il s'est avéré que le corps de la langue a un effet significatif sur F2 de toutes les voyelles, sur F1 de [i e œ ɔ] et sur F3 de [i e ɛ a ɔ œ]: il est considéré comme le geste le plus "audible" qui contrôle la cible acoustique des voyelles. Les travaux de Delattre (1971) et de Ghazeli (1977) avaient souligné l'effet du mouvement horizontal d'avancement et de rétraction du corps de la langue en arabe sur les structures formantiques des voyelles adjacentes. Ghazeli écrit à ce propos que : "the horizontal displacement of the tongue affects the vocal tract cavities in a way that is manifested in the transition to and away from the adjacent vowels" (Ghazeli 1977, p. 76).

Nous proposons ici la figure 1 pour montrer les différents lieux d'articulation des consonnes de l'ASM, corrélés au mouvement du corps de la langue. Il faut souligner que nous n'avons pas représenter sur cette figure le fait que les emphatiques sont réalisées avec une double constriction (alvéolaire et pharyngale), ainsi, le fait que Le voile du palais est abaissé pour les consonnes nasales /m/ et /n/.

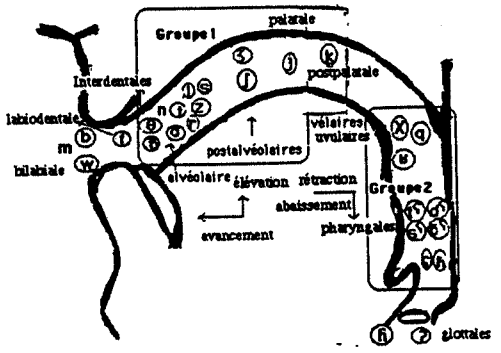


Figure1. Les différents lieux d'articulation des consonnes de l'ASM en fonction du mouvement du corps de la langue.

Le groupe 1 des consonnes linguales réalisées avec l'avancement et élévation du corps de la langue vers la partie antérieure du conduit vocal.
Le groupe 2 des consonnes linguales réalisées avec la rétraction et l'abaissement du corps de la langue.

A partir des faits mentionnés ci-dessus, il nous a semblé intéressant d'effectuer une analyse de l'écart entre F1 et F2 des voyelles de l'ASM où elles sont précédées des différentes consonnes linguales en structure CV pour évaluer l'effet du mouvement d'avancement et de retraction du corps de la langue sur les structures formantiques des voyelles adjacentes de l'ASM.
Notre hypothèse est que les cibles acoustiques des voyelles de l'ASM varient considérablement en contact des consonnes linguales selon leurs différents lieux d'articulation.
À noter que cette étude s'inscrit dans la série des travaux sur les variations de timbre des voyelles de l'ASM: Ghazeli 1979, Rajouani 1987, Znagui 1992, Znagui & Boudelaa 1993.

1. PROTOCOLE EXPÉRIMENTAL

Six locuteurs adultes maghrébins (2 marocains, 2 algériens et 2 tunisiens) ont été retenu pour les deux expériences de l'étude. Ils ont lu le corpus avec une voix normale. L'enregistrement a été effectué dans une chambre insonorisée sur un Revox A 72 à la vitesse de 7,5 cm/sec à l'aide d'un microphone professionnel ECM 265 situé à 30 cm de la bouche du locuteur. Une première analyse spectrographique, à bande large, a été effectuée, au Laboratoire de Phonétique de Paris VII, par le logiciel Macspeechlab, l'échantillonnage étant fixé à 10 kHz avec une dynamique de 12 bits. Nous avons également procédé à une analyse par le logiciel Unice, Au Laboratoire de Phonétique de Paris III, avec une dynamique de 16 bits. Les mesures spectrographiques des voyelles sont faites au milieu de la partie stable de F1 et F2 (cible acoustique).

1.2. PREMIÈRE EXPÉRIENCE

Dans cette première expérience, 392 logatomes ont été soumis à l'analyse spectrographique. Les 6 locuteurs ont répété 3 fois chaque contexte de 20 consonnes linguales, et 2 consonnes glottales. La voyelle /ə/ est analysée en structure VCV où C est l'une des consonnes linguales de l'ASM : /ð θ t d s z l ʃ ʒ j k χ q ʁ ʕ ħ tˤ sˤ ðˤ dˤ/. Le choix de la voyelle /ə/ est motivé par le fait que, dans sa production, le conduit vocal est en position la plus neutre selon Chiba et kajiama (1958) et qui correspond approximativement aux fréquences formantiques suivantes: F1 = 500 Hz, F2 = 1500 Hz, F3 = 3500 Hz. Nous estimons que les différents lieux d'articulation des consonnes linguales (interdentales, alvéolaires, posalvéolaire, palatale, pospalatale, vélaire-uvulaire, pharyngale et pharyngalisé)

constituent des régions de constrictions (linguales) successives dans le conduit vocal neutre représenté par la voyelle /ə/. Ce qui nous permet d'examiner les effets de la coarticulation linguale sur les trajectoires formantiques de cette voyelle en VCV.

Nous avons choisi la méthode LPC (codage par prédiction linéaire), qui a l'avantage de déterminer les coefficients de prédiction en minimisant l'erreur entre l'entrée et le signal prédit en sortie. Elle permet aussi d'observer facilement les directions des transitions formantiques.

Nous avons pu ainsi observer l'effet de coarticulation des consonnes linguales sur les trajectoires formantiques de la voyelle neutre /ə/, et comparer nos résultats avec ceux de Delattre (1967), de Mrayati et al.(1988) et de Dakkak et al. (1992), même si la voyelle de l'étude /ə/ est différente de /ɛ/. Nos résultats sont représentés sur la figure 2 ci-dessous.

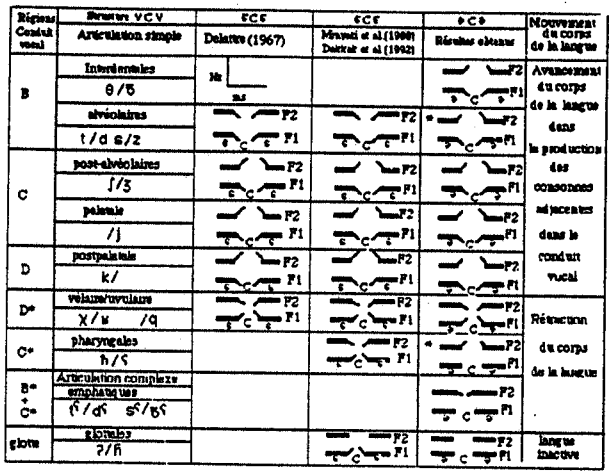


Figure 2. Trajectoires de F1 et F2 pour la voyelle /ə/ avec les différents lieux d'articulation des consonnes de l'ASM, en structure CV, et en fonction du mouvement du corps de la langue.

Sur la figure, nous n'avons pas montré la barre de voisement de la voyelle, ni les structures spectrales des consonnes. Les consonnes voisées sont indiquées à droite de la barre oblique (/) et les consonnes sourdes à gauche.

Notre analyse spectrographique montre des régularités dans les directions des transitions de F1 et F2 de la voyelle /ə/ en contact des différents lieux d'articulation des consonnes linguales comme suit :

1- Un groupe homogène:

Au contact des consonnes linguales antérieures (interdentales, alvéolaires, postalvéolaires, palatales et postpalatales) : les transitions de F1 sont négatives (descendantes) et les transitions de F2 sont positives (montantes). Ces trajectoires sont conformes aux prédictions que l'avancement du corps de la langue fait baisser les fréquences de F1 et augmente les fréquences de F2 (Delattre (1971).

2- Un groupe moins homogène:

a) Au contact des consonnes vélaires-uvulaires, les transitions de F1 sont positives et celles de F2 sont négatives. Ces trajectoires sont conformes aux prédictions qui montrent que le recul du corps de la langue provoque un resserement de la région vélo-pharynx ce qui provoque une élévation fréquentielle de F1 et l'abaissement de F2.

b) Au contact des consonnes pharyngales, la direction des trajectoires semble ne pas se conformer aux prédictions qui correspondent aux transitions positives pour F1 et négatives pour F2 (Dakkak et al.(1992). En effet, nous constatons, chez nos locuteurs maghrébins, des transitions légèrement négatives ou stables pour F1 et légèrement positives ou stables pour F2. Il reste souligner que l'écart entre F1 et F2 de la voyelle /ə/ est plus réduit dans leur partie stable par rapport à celui cette voyelle réalisé en contact des consonnes linguales antérieures du premier groupe.

c) Au contact des consonnes pharyngalisées réalisées avec une double constriction (alvéolaire et pharyngale) dans le conduit vocal, les transitions de /ə/ sont stables pour F1 et stables ou négatives pour F2. On observe un rapprochement très net des formants.

Il ressort de cette première expérience que les trajectoires de F1 et F2 de la voyelle /ə/ au contact des consonnes postérieures sont moins homogènes que celles qui sont au contact des consonnes antérieures. Toutefois l'écart entre F1 et F2 dans la partie stable des voyelles reste un facteur de différenciation rigoureux entre les deux groupes cités ci-dessus. C'est ce que l'on va essayer de démontrer dans la deuxième expérience.

1.3. DEUXIÈME EXPÉRIENCE

Dans cette expérience, nous avons élaboré un corpus de voyelles de l'ASM en structure CV dans des mots réels (2376 items) où V est, soit une voyelle brève inaccentuée pour évaluer son degré de perméabilité au contexte, soit une voyelle longue accentuée pour y tester son degré de résistance à la coarticulation linguale. Le choix de la structure CV (syllabe ouverte) est motivé par le fait que, dans le système linguistique de l'ASM, le mot ne commence jamais par une voyelle. De ce fait, la voyelle en contexte isolé n'a pas de réalité psychologique dans cette langue.

Les 6 voyelles de l'ASM /a i u, a: i: u:/ sont donc précédées des consonnes linguales /ð θ t d l s z ʃ ʒ j k χ q ʁ ʕ ħ tˤ sˤ ðˤ dˤ/ et insérées dans une phrase porteuse : "qa:la /C1V1C2V':2CVC/ marratan " (il a dit [...] une fois) où C1 est l'une des 22 consonnes linguales et V1 est la voyelle de l'étude. C2 est la consonne labiale (ou glottale dans certains cas) en position fixe qui selon Amerman et Daniloff (1977) "would offer minimal interference with coarticulation of tongue position" [Amerman et Daniloff (1977), p.107]. Nous avons également introduit d'autres mots dans le corpus pour détourner l'attention des locuteurs de l'objet de notre travail.

Nous avons choisi les valeurs formantiques des voyelles précédées des consonnes glottales comme référence, du fait qu'elles ne sont pas réalisées avec la langue dans le but de mettre en évidence l'effet du corps de la langue dans les deux cas : actif et passif. De même, Nous n'avons pas pris en considération les valeurs formantiques des voyelles précédées des consonnes labiales bien qu'elles ne soient pas réalisées avec la langue comme les glottales, puisqu'il y a les effets de la labialité qui font baisser les fréquences de tous les formants.

2. ANALYSE ET RESULTATS

L'analyse spectrographique des voyelles /a i u a: i: u:/ montre des régularités dans les variations formantiques de F1 et F2 (mesures effectuées dans les parties stables des deux formants) en CV. Nous utilisons ici le terme vocoïde pour préciser qu'il s'agit des réalisations d'un même phonème vocalique (niveau phonétique). Nous avons constaté que l'écart entre F1 et F2 des vocoïdes précédés des consonnes linguales antérieures /ð θ t d l s z ʃ ʒ j k/ (groupe GV1) est nettement plus élevé que celui des vocoïdes précédés des consonnes linguales postérieures /χ q ʁ ʕ ħ tˤ sˤ ðˤ dˤ/ (groupe GV2). GVG est le goupe de vocoïdes précédés des consonnes glottales /ɦ ʔ/. Voir tableau 1:

moyen/Hz	F1 /i/	F2 /i/	F2-F1/i/
GV1	302	2202	1900
GVG	353	2231	1878
GV2	348	2095	1747
moyen/Hz	F1 /a/	F2 /a/	F2-F1/a/
GV1	470	1605	1135
GVG	580	1594	1014
GV2	661	1387	725
moyen/Hz	F1 /u/	F2 /u/	F2-F1/u/
GV1	335	1152	817
GVG	381	980	599
GV2	421	1008	587
moyen/Hz	F1 /i:/	F2 /i:/	F2-F1/i:/
GV1	301	2338	2037
GVG	340	2205	1865
GV2	374	2170	1796
moyen/Hz	F1 /a:/	F1 /a:/	F2-F1/a:/
GV1	496	1706	1210
GVG	520	1594	1074
GV2	584	1335	751
moyen/Hz	F1 /u:/	F1 /u:/	F2-F1/u:/
GV1	339	1142	803
GVG	360	1066	706
GV2	387	975	588

Tableau 1 : Valeurs moyennes de F2 et F1 des différents groupes de vocoïdes brèves et longues de l'ASM avec leurs différences.

Nous avons appliqué le test-t apparié unidirectionnel sur les valeurs moyennes de la différence entre F2 et F1 des deux groupes de vocoïdes (GV1 et GV2) qui s'est révélé significatif : $t(5) = 5{,}76$, $p<0{,}002$). Les figures 3-4 montrent l'écart entre F1 et F2 des deux groupes de vocoïdes de l'ASM.

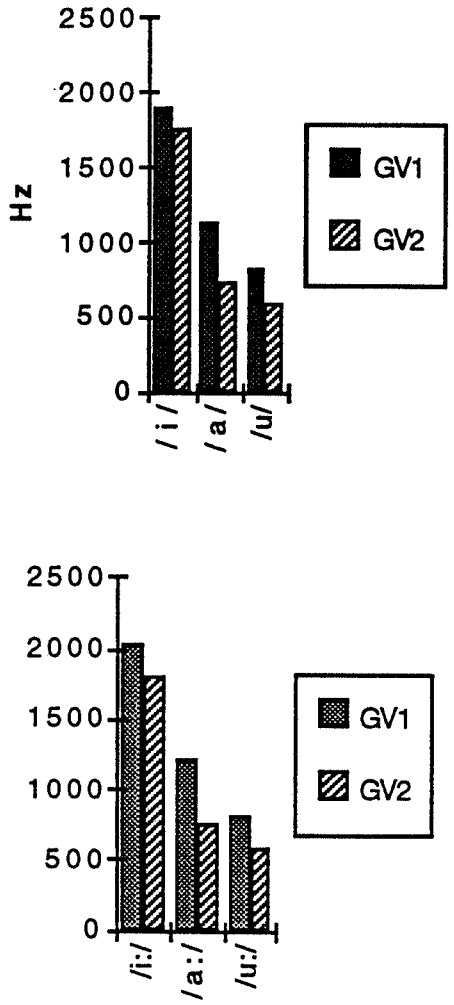


Figure 5. L'écart entre F1 et F2 des deux groupes de vocoïdes brefs et longs de l'ASM.

Les résultats de notre analyse montrent que le système vocalique de l'ASM peut être décrit du point de vue du mouvement de la langue en deux groupes de vocoïdes : un premier groupe (G1) antériorisé dans son lieu d'articulation au contact des consonnes / ð θ t d n s z ʃ ʒ j k / réalisées avec un déplacement horizontal du corps de la langue en avant du conduit vocal : [i̟ i̟ː a̟ a̟ː u̟ u̟ː], corrélé à un éloigement de F2 et F1 et une fermeture de l'aperture. Un second groupe (G2) précédé de consonnes /χ q ʁ ʕ ħ tˤ sˤ ðˤ dˤ/ réalisées avec une rétraction du corps de la langue vers la partie arrière du conduit vocal : [i̠ i̠ː a̠ a̠ː u̠ u̠ː], et corrélé à un rapprochement de F2 et F1 et à une ouverture de l'aperture. voir la figure 6.

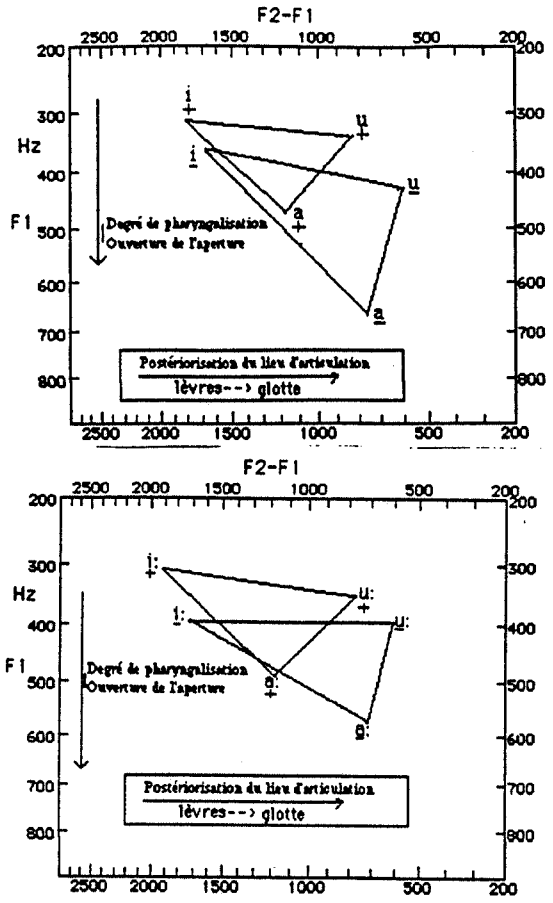


Figure 6. Trapèzes vocaliques des deux groupes de vocoïdes de l'ASM

Quant à évaluer directement les conséquences de ces résultats illustrés sur les trapèzes vocaliques ci-dessus, il ressort que :

Pour le Groupe 1 de vocoïdes antériorisés, les fréquences formantiques de F1 et F2 correspondent aux timbres suivants:

i ==> [i], a ==>[ɛ ae] , u ==>[u]

i: ==>[e:], a: ==>[ɛ: ae:], u: ==>[u:]

Pour le groupe 2 de vocoïdes postériorisés, les fréquences formantiques de F1 et F2 correspondent aux timbres suivants:

i ==>[e], a ==>[a ɑ], u ==>[o]

i: ==>[e:], a: ==>[a: ɑ:], u: ==>[o:]

En fait, ces résultats mettent l'accent sur le rôle important de la coarticulation linguale dans les modifications des cibles acoustiques des voyelles adjacentes brèves inaccentuées et longues accentuées de l'ASM. En soulignant toutefois que le degré de la coarticulation varie suivant que la voyelle est fermée comme /i/ et /u/ ou ouverte /a/. En effet, nous constatons que les écarts entre F2 et F1 des vocoïde antériorisés [i] et [u] et des vocoïdes postériorisés [i] et [u] ne sont pas nettement marqués par rapport à l'écart entre F2 et F1 du vocoïde antériorisé [a] et le vocoïde postériorisé [a]. (voir graphique 7). Nous pouvons dire que la voyelle ouverte /a/ en position accentuée et inaccentuée est la plus vulnérable à la coarticulation linguale. De même, Il semble que la voyelle de l'ASM qu'elle soit en position accentuée ou longue n'arrive pas à freiner les effets de la coarticulation linguale. Il reste à souligner que la voyelle brève /a/ et la voyelle /a:/ présentent une grande variabilté acoustique qui correspond aux timbres de [ɛ ae a ɑ] corrélées à une grande latitude articulatoire. Cependant, à partir de nos résultats sur la coarticulation linguale, nous pouvons affirmer que cette dispersion acoustique présente deux allophones [ae] et [a] qui dépendent du mouvement d'avancement et de rétraction du corps de la langue. Nous proposons la figure 7 qui montre le degré de coarticulation linguale en fonction de l'aperture de la voyelle.

Nature des vocoïdes	moyenne F2-F1
G1 Vocoïdes fermés non arrondies [i]- [ɪ] et [i:] -[ɪ:]	394 Hz
G2 Vocoïdes fermés arrondies [u]-[ʊ] et [u:]-[ʊ:]	445 Hz
G3 Vocoïdes ouverts [a]-[ɑ] et [a:]-[ɑ:]	869 Hz

Figure 7. Degré de coarticulation linguale illustré par l'écart entre F1 et F2 des différents groupes de vocoïdes brefs et longs de l'ASM.

En résumé, nous proposons une classification acoustico-phonétique binaires pour les vocoïdes brèves et longues de l'ASM. Voir figure 8.

Mouvement de la langue	lieu d'articulation des consonnes adjacentes	Les vocoïdes de l'ASM en structure CV	
		i a u ɪ ɑ ʊ	i a u ɪ ɑ ʊ
Le corps de la langue est avancé	Interdentale Alvéolaire Post alvéolaire palatale postpalatale	F2 et F1 sont moins rapprochés - Ouverts + Antériorisés	
Le corps de la langue est retracté	Vélaire-uvulaire Pharyngale Interdentale pharyngalisé Alvéolaire		F2 et F1 sont plus rapprochés + ouverts + Postériorisés

Figure 8. Classification binaire acoustico-phonétique des vocoïdes brefs et longs de l'ASM.

CONCLUSION

Nous avons adopté une approche acoustico-phonétique dans l'analyse des voyelles de l'ASM précédées des consonnes linguales de différents lieux d'articulation. Cette approche nous a fourni des résultats interessants sur la coarticulation linguale gauche-droite en CV. Elle confirme notre hypothèse selon laquelle l'écart entre F1 et F2 des voyelles est fortement modifié par l'entourage des consonnes

linguales. Il en résulte que les consonnes linguales réalisées, soit dans la partie antérieure du conduit vocal, soit dans sa partie postérieure imposent des cibles acoustiques bien déterminées aux voyelles adjacentes et à des degrés différents en fonction de l'aperture fermée ou ouverte de ces voyelles. Chaque phonème vocalique des 6 voyelles de l'ASM est réalisé en deux allophones :

- un allophone antériorisé dans son lieu d'articulation et corrélé sur le plan acoustique à un éloignement entre F1 et F2
- Un allophone postériorisé dans son lieu d'articulation et corrélé sur le plan acoustique à un rapprochement de F1 et F2.

Enfin, ce travail montre que les conséquences du mouvement horizontal du corps de la langue dans la production des consonnes permettent de prédire des régularités dans les variations formantiques des voyelles de l'ASM dont on propose de tester la validité sur le plan perceptif dans nos travaux ultérieurs. Nul doute qu'une telle étude est nécessaire pour la description de la phonétique de l'ASM, et qui comme nous le pensons, répond à des besoins précis en ce qui concerne le contrôle des paramètres de commande articulatoire et acoustique en synthèse et reconnaissance automatique de la parole.

REMERCIEMENTS

Nous tenons à remercier Mr. Gsell, S. Maeda, N. Clements, S. Ghazeli, G. Bohas, D. Rostolland, B. Gautheron, Boulakia, G., Mme J. Vaissière et Mme A. Rialland pour leurs conseils, leurs encouragements et leurs précieux commentaires lors de l'élaboration de cet article.

RÉFÉRENCES

- Amerman, J. D. & Daniloff, R. G . (1977) , Aspects of Lingual Coarticulation, Journal of Phonetics, 5, pp. 107-113.
- Boé, L.J.& Perrier P.& Morris A. (1992) Une prédiction de l'audibilité des gestes articulatoires à partir d'une modélisation articulatoire, 19 JEP, pp.151-157.
- Chiba, T. & Kajiama, M. (1958) The vowel its Nature and Structure, Ed. Phonetic Society of Japan, Tokyo.
- Chistovitch, L.A.& Lublinskaya, V.V. (1979), The center of gravity effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli, Hearing Research,1., pp.185-195.
- Dakkak, O. & Mrayati, M. & Carré, R. (1992) Transitions formantiques correspondant à des constrictions réalisées dans la partie arrière du conduit vocal, 19 JEP, pp.167-170.
- Delattre, P. (1971) Pharyngeal features in the consonants of Arabic, German, French and American English, Phonetica 21, pp.129-155.
- Delattre; P. (1967) Des indices acoustiques aux traits pertinents, 6 th Int. Congr. Phonetic. S. pp.35-46.
- Fant, G. (1960) The Acoustic Theory of Speech Production, The Hague, Mouton.
- Ghazeli, S. (1977) Back Consonants and Backing Articulation in Arabic Unpublished Phd. Dissertation, University of Texas At Austin.
- Ghazeli, S. (1979) Du Statut des voyelles en Arabe, Analyse Théorie 2/3, pp. 199-219.
- Lindblom, B. & Sundberg, J. (1971) Acoustical consequences of lip, tongue, jaw and larynx movements, J.A.S.A., 50, pp. 1166-1179.
- MacNeilage, P. F.& Declerk, J. L. (1969) On the Motor Control of Coarticulation in CVC Monosyllables, J.A.S.A. 45, pp. 1217-1233.
- Merlmestein, P., (1973) Articulatory Model For The Study of Speech Production, JASA, 53, (4), pp. 1070-1082.
- Mrayati, M. & Carré, R. & Guérin, B. (1988) Distinctive regions and mode: A new theory of speech production, Speech Communication, 7, pp.257-286.
- Öhman, S. (1966) Coarticulation in VCV utterances: Spectrographic measurements: J.A.S.A., 39, pp. 151-168.
- Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study, MIT Press, Cambridge.
- Rajouani, A. & Najim, M. & Chiadmi, D. (1987) Synthesis of pharyngealisation features in Arabic, Speech Communication, 6, pp.261-268.
- Recasens, D. (1991), An Electropalatographic and Acoustic Study of Consonant to Vowel Coarticulation, Journal of Phonetics 19, pp.177-192.
Schouten, M.E.H.& Pols, L.C.W. (1979) Vowel segments in consonantal contexts: a spectral study of coarticulation, Journal of phonetic, 7, pp. 1-23.
- Stevens, K.N. (1989) On the Quantal Nature of Speech, Journal of Phonetics, 17, pp. 3-45.
- Znagui, I., & Boudelaa, S. (1993) Towards An Acoustic-Phonetic Classification Of Modern Standard Arabic Vowels, EuroSpeech, pp. 1425-1427.
- Znagui. I. (1992) Etude formantique des voyelles de l'arabe standard, 19 JEP, Bruxelles, pp. 493-497.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O3

TRAITEMENT DU SIGNAL

ADAPTATION AUX CONDITIONS DES LIGNES TÉLÉPHONIQUES D'UN SYSTÈME DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

C. MOKBEL, P. PACHÈS-LEAL, J. MONNÉ & D. JOUVET

FRANCE TÉLÉCOM, CNET-LAA/TSS/RCP,
2 ROUTE DE TRÉGASTEL, BP 40, 22300 LANNION CEDEX

Résumé

Nous traitons de la compensation des effets des lignes téléphoniques sur le signal de parole pour l'amélioration des performances de reconnaissance. Nous montrons par les expériences que le canal téléphonique agit comme un filtre linéaire qui varie lentement au cours d'un appel. Ceci se traduit par des biais entre les espaces acoustiques des différents appels. Nous démontrons que si, sur un horizon temporel donné, l'effet du canal est constant et le signal propre est noncohérent, le cepstre du canal est la moyenne des vecteurs cepstraux du signal observé. Ceci justifie les techniques classiques comme la soustraction cepstrale ou le filtrage passe-haut des trajectoires cepstrales (connu comme RASTA [2]). Les améliorations des performances de reconnaissance avec ces techniques sont statistiquement significatives. Les expériences de reconnaissance sont menées en mode mots isolés et indépendant du locuteur. Nous avons essayé d'utiliser des techniques neuronales pour réduire les biais dus aux lignes téléphoniques. Les résultats obtenus montrent des performances inférieures aux techniques classiques.

Mots-clés: *reconnaissance, déconvolution, effets de lignes, cepstre à long-terme, réseau de neurones.*

I. INTRODUCTION

Dès leur sortie des laboratoires, les systèmes de reconnaissance automatique de la parole (RAP) se sont heurtés au problème de l'environnement. En effet, l'acquisition du signal ne se fait pas dans des conditions idéales. Rien que le changement du microphone peut détériorer d'une façon dramatique les performances de la reconnaissance. On distingue généralement deux composantes perturbatrices dans le signal observable à l'entrée d'un système de reconnaissance. La première est due au bruit ambiant et est considérée comme additive au signal propre de la parole. La seconde est convoluée au signal propre et dépend du canal de transmission véhiculant l'information acoustique du locuteur jusqu'au système de reconnaissance, y compris le microphone.

La reconnaissance à travers le réseau téléphonique, en particulier pour les serveurs vocaux interactifs, est une application potentielle de la RAP. Nos travaux se situent dans ce cadre. Les études présentées dans cet article traitent du problème de la compensation des effets des lignes téléphoniques sur le signal de parole dans le but d'une meilleure reconnaissance de ce signal. On considère que les effets de lignes se manifestent dans le signal observable comme une composante à caractère essentiellement convolutif. Ainsi, le problème se réduit à un problème classique de déconvolution.

Pour augmenter la robustesse des systèmes de reconnaissance aux variations des conditions des lignes téléphoniques, plusieurs approches sont proposées dans la littérature. Elles sont basées sur un filtrage des trajectoires cepstrales ou log-spectrales où la perturbation convolutive introduite par le canal se transforme en une composante additive qui introduit un biais variable selon les appels. Vu les caractéristiques lentement variables des lignes téléphoniques, ces biais sont généralement considérés comme presque constants par appel.

Tout le long de ce travail des bases de données téléphoniques seront utilisées. Ainsi, nous commençons par décrire l'ensemble de ces bases dans la section II.

Des mesures sont effectuées sur une base de données téléphoniques mettant en évidence les effets des lignes et leur caractère convolutif. La section III présente ces mesures ainsi qu'une formulation du problème.

Malheureusement, dans l'application envisagée nous disposons d'une seule observation délivrée par le microphone du combiné téléphonique. Nous ne pouvons pas donc utiliser des techniques de déconvolution basées sur le filtrage spatial, le traitement d'antenne ou la déconvolution aveugle.

Dans une première approche neuronale, un modèle de production du signal observable est introduit. Ce modèle est indispensable pour la définition de la stratégie à suivre. C'est l'objet de la section IV.

Dans la section V, nous démontrons un théorème qui lie la moyenne des vecteurs cepstraux à l'effet des lignes téléphoniques. Ce théorème justifie le fait que la suppression de la partie continue des vecteurs cepstraux réduit les effets des lignes téléphoniques. Ainsi, nous expérimentons la soustraction cepstrale qui consiste à moyenner les vecteurs cepstraux et à soustraire la moyenne trouvée. Une approche équivalente consiste à

appliquer sur les trajectoires cepstrales un filtrage passe-haut avec un temps de réponse suffisamment élevé [2][4]. Ces deux approches sont testées en reconnaissance sur plusieurs corpus, de laboratoire et d'exploitation. Elles donnent pleine satisfaction en améliorant les performances des systèmes de reconnaissance d'une manière statistiquement significative.

Finalement dans la section VI, nous reportons un ensemble de conclusions, et identifions les points à approfondir.

II. BASES DE DONNÉES UTILISÉES

Plusieurs bases de données sont utilisées dans nos expériences. Elles sont toutes collectées à travers le réseau téléphonique. Nous distinguons deux types: les bases de laboratoire et celles d'exploitation. Pour les bases de laboratoire, on demande aux locuteurs de répéter des listes prédéfinies du vocabulaire. En revanche, les données d'exploitation sont formées de vrais appels à des serveurs mis en exploitation.

Les corpus utilisés dans nos expériences sont: les chiffres isolés (taille 7200 mots), le Trégor (36 mots isolés) (taille 25500 mots), les nombres à 2 chiffres (de 00 à 99) (taille 15500 nombres), le dictionnaire de 500 mots et les Baladins (26 mots) (taille 23000 mots). Les trois premiers, des données de laboratoire, sont collectés par environ 800 locuteurs chacun. Les locuteurs appellent de différentes régions de France. Les mots détectés automatiquement sont validés à l'écoute. Le dictionnaire de 500 mots est aussi de laboratoire et est enregistré par 10 locuteurs à raison de 3 répétitions par locuteur. Le corpus des Baladins est formé de deux parties: une de laboratoire (500 locuteurs) et une d'exploitation (1500 appels).

III. DÉFINITION DU PROBLÈME

La phase d'acquisition du signal de parole influe sur les performances de la reconnaissance automatique de ce signal. Nous nous intéressons à une reconnaissance à travers le réseau téléphonique où la phase d'acquisition se schématise comme dans la figure 1.

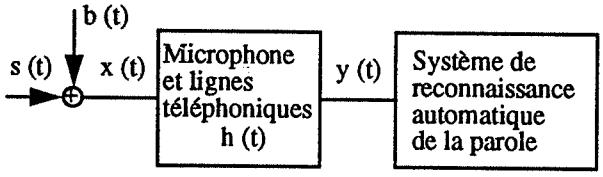


Figure 1: *Acquisition d'un signal téléphonique.*

Une fois la liaison téléphonique avec le serveur vocal établie, l'utilisateur prononce des mots ou des phrases. Ses réalisations acoustiques s (t) sont transformées par le microphone du combiné en signal électrique. Malheureusement, le microphone capte aussi le bruit ambiant b (t). Ces deux signaux sont ainsi additionnés pour donner un signal x (t) qui traverse la liaison téléphonique pour arriver au serveur vocal. Cette dernière se caractérise généralement par un filtre linéaire que l'on nomme h (t). Ainsi, le signal observé en entrée du système de reconnaissance y (t) peut s'écrire en fonction de toutes ces composantes:

$$y(t) = x(t) \otimes h(t) = [s(t) + b(t)] \otimes h(t) \qquad (1)$$

L'équation (1) est une formulation du problème schématisé dans la figure 1. Le bruit ambiant b (t) s'additionne au signal propre s (t) pour produire x (t) qui se convolue à h (t) en traversant le canal téléphonique, donnant y (t) comme observation à l'entrée du système de reconnaissance.

Dans le cadre des serveurs vocaux interactifs qui offrent des services d'information à travers le réseau téléphonique, les environnements des utilisateurs sont en général peu bruyants et les mots à reconnaître sont prononcés près du microphone du combiné. L'effet du bruit ambiant additif b (t) est donc négligeable et le rapport signal/bruit additif est assez élevé. Cette hypothèse permet de simplifier l'équation (1) qui devient:

$$y(t) = s(t) \otimes h(t) \qquad (2)$$

Dans l'équation (2), seuls les effets des lignes téléphoniques perturbent le signal de parole. Réduire cette perturbation est un problème de déconvolution.

La nature des communications téléphoniques induit une autre hypothèse. En effet, une fois la liaison téléphonique établie, ses caractéristiques restent constantes ou varient lentement au cours du temps.

Pour une vision plus claire, la composante perturbatrice h (t) introduite dans le signal temporel doit être projetée dans l'espace de représentation où s'effectue le décodage acoustique. L'espace de représentation cepstrale est le plus couramment utilisé. Le système de reconnaissance du CNET utilise une analyse MFCC. Les vecteurs acoustiques sont de dimension 27 et sont formés du logarithme de l'énergie et des 8 premiers coefficients cepstraux auxquels on ajoute leurs dérivées premières et secondes. Les MFCC, coefficients cepstraux, sont obtenus par une transformation de Fourier inverse, du logarithme des énergies en sortie d'un banc de filtres répartis sur une échelle perceptive MEL. Si on calcule les densités spectrales des deux termes de l'équation 2, nous obtenons:

$$\Gamma_y(f) = \Gamma_s(f) \cdot \Gamma_h(f) \qquad (3)$$

En appliquant la fonction logarithme et par transformation de Fourier inverse, nous obtenons les coefficients cepstraux. Soit:

$$C_y(\tau) = C_s(\tau) + C_h(\tau) \qquad (4)$$

Les effets des lignes téléphoniques se manifestent dans l'équation (4) par $C_h(\tau)$, un biais dans l'espace cepstral. Ce biais presque constant pour un appel donné, varie d'un appel à un autre. Cette variation a pour conséquence l'augmentation des variabilités dans l'espace

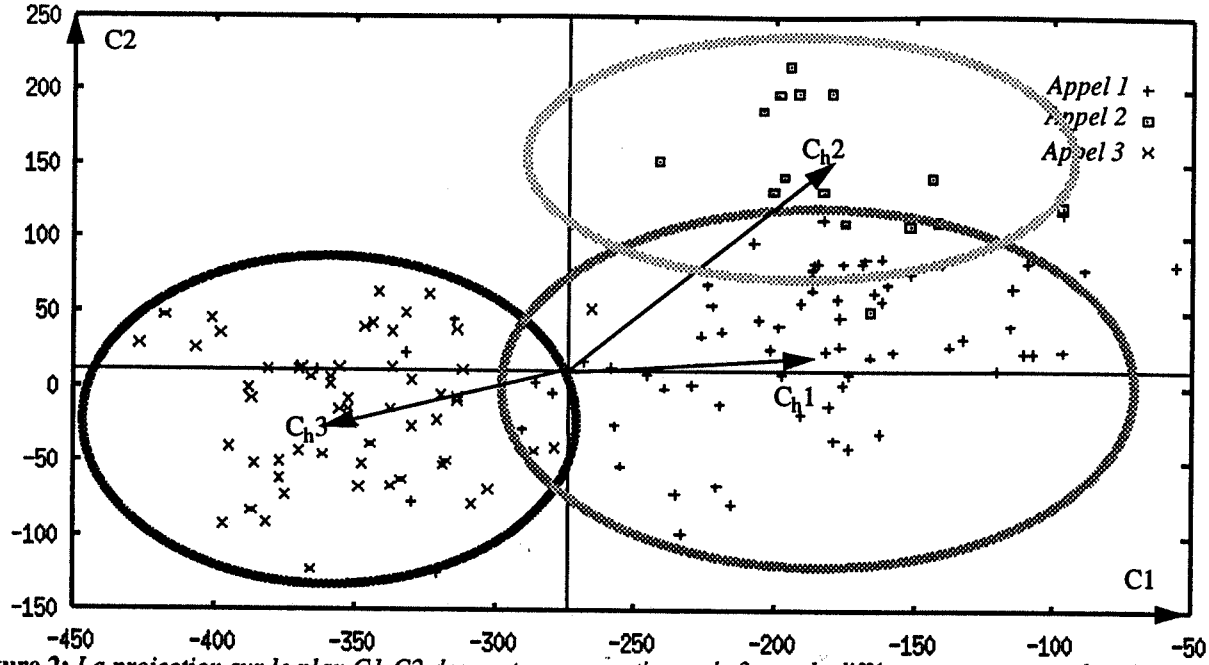


Figure 2: *La projection sur le plan C1-C2 des vecteurs acoustiques de 3 appels différents et correspondant à une gaussienne de silence. On voit clairement les translations dues aux effets des lignes téléphoniques.*

cepstral et la diminution de son caractère discriminant.

Afin de valider les hypothèses et formulations précédentes, des mesures sont effectuées sur la base de données des chiffres. Au travers de ces mesures nous cherchons à montrer qu'***un bruit convolutif existe et varie lentement au cours d'un appel.*** Cette perturbation se traduit par ***un vecteur de translation dans l'espace cepstral presque constant par appel, mais variable d'un appel à un autre. Ces biais sont mal traités par un modèle de Markov.***

Nous nous contentons de présenter une des mesures effectuées. Des modèles de Markov sont appris pour modéliser les chiffres. Ces modèles comprennent un modèle commun du silence de début et de fin des enregistrements permettant de détecter implicitement les frontières des chiffres. En alignant les chiffres de trois appels sur leurs modèles respectifs, nous collectons les vecteurs correspondant à une gaussienne du modèle de silence. Ces vecteurs sont projetés sur le plan C1-C2 des deux premiers coefficients cepstraux. Cette projection est montrée dans la figure 2. Sur cette figure, nous avons aussi tracé les projections des cepstres à long-terme des trois appels: C_h1, C_h2 et C_h3. Il est clair que les trois ensembles de points correspondent à une gaussienne assez compacte mais translatée par les effets des lignes téléphoniques dans trois directions différentes. Les vecteurs acoustiques translatés occupent dans le modèle Markovien une gaussienne plus large que celle d'origine. La soustraction des vecteurs C_h1, C_h2 et C_h3 des vecteurs acoustiques des appels correspondants, réduit la variabilité due aux lignes téléphoniques et augmente le caractère discriminant de l'espace acoustique.

La figure 2 et les remarques relatives permettent de valider expérimentalement la formulation des effets des lignes décrites par les équations (3) et (4).

IV. DÉCONVOLUTION PAR RÉSEAU DE NEURONES

A l'origine, l'utilisation des réseaux de neurones est guidée par l'intuition décrite dans la figure 3.

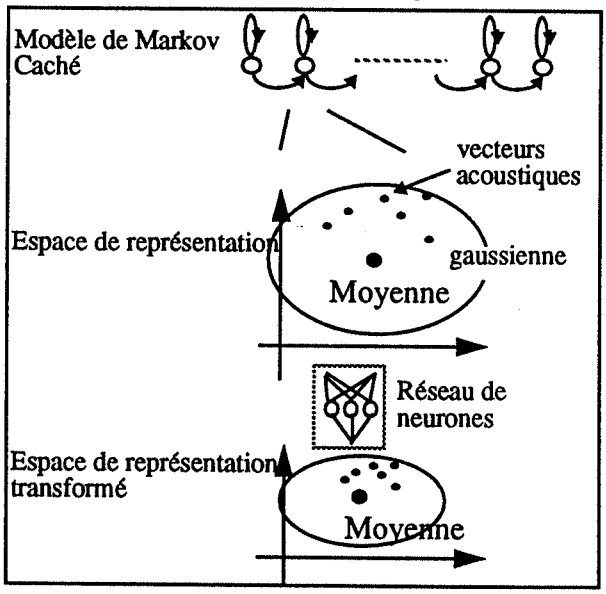


Figure 3: *Intuition à l'origine de l'utilisation du connexionisme (réduction de la dispersion des vecteurs).*

La recherche d'une transformation rapprochant les vecteurs acoustiques de leurs moyennes correspondantes réduirait les variances des gaussiennes et augmenterait la discrimination entre les différentes formes améliorant ainsi les performances en reconnaissance. Afin de compléter l'intuition décrite précédemment, nous proposons un modèle de production (figure 4) des vecteurs acoustiques correspondant à la réalisation d'un mot ou d'une phrase du vocabulaire. Ce modèle suppose que lors de la production d'une forme du vocabulaire de référence,

l'utilisateur émet une suite de vecteurs de référence selon un processus de Markov. A ces vecteurs de référence C_r, vient s'ajouter une variable vectorielle aléatoire perturbatrice C_l qui représente la variabilité intrinsèque au locuteur. Sous l'hypothèse d'un bruit additif négligeable devant le signal de parole, la perturbation extrinsèque à la parole se présente comme une variable aléatoire C_h qui s'ajoute au signal propre pour donner les vecteurs observables C_y. Notons que les vecteurs représentés dans ce modèle sont des vecteurs cepstraux.

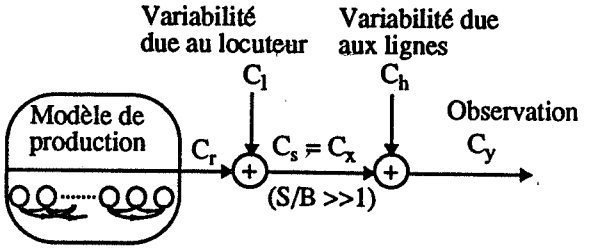


Figure 4: *Modèle de production des observations.*

Suivant ce modèle, l'amélioration des performances de la RAP se fait en cherchant une transformation universelle rapprochant les vecteurs observés C_y des vecteurs de référence C_r. Une telle transformation peut être identifiée et estimée par un perceptron multicouche. Reste à définir les vecteurs de référence C_r. Pour ce faire, nous alignons les vecteurs acoustiques de l'ensemble d'apprentissage sur les modèles de Markov correspondants. Cet alignement délivre pour chaque vecteur acoustique le vecteur moyenne de la gaussienne correspondant. Ces moyennes de gaussiennes sont considérées en une première approximation comme les vecteurs de référence C_r. Ainsi, un perceptron multicouche est appris pour rapprocher les vecteurs acoustiques des moyennes des gaussiennes correspondantes dans les modèles de Markov. Ensuite les vecteurs acoustiques sont tous transformés par le perceptron appris. Les expériences de reconnaissance sont menées sur les vecteurs transformés.

Les premières expériences ont montré qu'une simple transformation des vecteurs acoustiques n'améliore pas leur caractère discriminant. Il faut donc procurer au réseau des informations supplémentaires. Ainsi le cepstre à long-terme (qui représente le canal téléphonique) sera donné en plus des vecteurs acoustiques en entrée du réseau qui cherchera de réduire les effets des lignes téléphoniques.

La fonction de coût à optimiser par le réseau étant extrêmement complexe, une bonne initialisation des poids des connexions est indispensable. Le réseau doit apprendre une transformation proche de l'identité (les vecteurs transformés ne doivent pas être très éloignés des vecteurs acoustiques de base). Afin de contourner les contraintes liées à une telle initialisation, nous demandons au réseau de calculer à sa sortie C_y-C_r au lieu de C_r. Finalement, comme les effets des lignes varient lentement, nous lissons la sortie du réseau sur 10 trames.

Des expériences de normalisation et de reconnaissance sont menées sur deux bases de données; le corpus des chiffres isolés et le corpus du Trégor. Le tableau 1 présente les taux obtenus après normalisation par réseau et compare aux résultats de base. Il est à noter que nous utilisons des modèles de Markov à 30 états pour modéliser chacun des mots du vocabulaire avec des modèles de silence de début et de fin de mot.

Tableau 1 : *Résultats de reconnaissance sans et avec déconvolution par réseau de neurones.*

Chiffres	taux d'erreur	*réduction du taux d'erreur*
Système de base	0,83%	
Seuil (amélioration significative à 95%)	**0,58%**	*30%*
Après normalisation par réseau	0,61%	*27%*
Trégor		
Système de base	0,64%	
Seuil (amélioration significative à 95%)	**0,52%**	*18%*
Après normalisation par réseau	0,62%	*3%*

Les résultats obtenus par normalisation en utilisant des réseaux de neurones montrent une amélioration timide par rapport aux résultats de base. Néanmoins, ils prouvent la possibilité d'utilisation des réseaux de neurones dans des schémas de déconvolution. La mise en oeuvre de telles techniques reste tout de même délicate.

V. SOUSTRACTION CEPSTRALE ET FILTRAGE PASSE-HAUT

Les travaux de recherche se sont multipliés ces dernières années pour réduire les effets des lignes téléphoniques dans le signal de parole observé à l'entrée d'un système de reconnaissance centralisé. La bibliographie riche dans ce domaine [2][4][1][5][3], conclut sur un effet convolutif des lignes téléphoniques variant lentement dans le temps. Un résultat pertinent de cette bibliographie est que la soustraction de la partie continue des vecteurs cepstraux (ou spectre logarithmique) estimée sur un horizon suffisamment large permet d'améliorer les performances d'un système de RAP. Cette réduction des effets des lignes téléphoniques peut se faire d'une manière explicite [5], estimer le cepstre à long-terme et le soustraire ensuite, ou implicite à travers un filtrage passe-haut appliqué sur les coefficients des vecteurs acoustiques comme dans la technique RASTA [2][4].

Des résultats positifs sont obtenus par cette appro-

che et les améliorations sont constantes dans tous les travaux publiés. Néanmoins, des mesures expérimentales sur des données téléphoniques comme celle du paragraphe III sont toujours utilisées pour justifier cette méthode. Cependant, on peut vérifier théoriquement le lien entre la moyenne des vecteurs cepstraux et les effets des lignes téléphoniques à partir d'hypothèses sur la cohérence du signal [6].

Théorème: *Sur l'horizon où l'effet des lignes téléphoniques se présente comme un filtre linéaire invariant dans le temps, la moyenne des vecteurs cepstraux représente le cepstre du canal si le signal propre est noncohérent.*

En effet, moyenner les vecteurs cepstraux revient à moyenner les logarithmes des spectres de puissance correspondants. Or la moyenne arithmétique des logarithmes des spectres de puissance est égale au logarithme de la moyenne géométrique des spectres de puissance.

$$\sum_{i=0}^{N-1} C_y^i \Leftrightarrow \sum_{i=0}^{N-1} \log\left[\Gamma_y^i(f)\right] = \log\left(\prod_{i=0}^{N-1} \Gamma_y^i(f)\right) \qquad (5)$$

Sous l'hypothèse que le canal téléphonique agit en filtre linéaire invariant dans le temps, nous réecrivons l'équation (3).

$$\Gamma_y^i(f) = \Gamma_s^i(f)\ |H(f)|^2 \qquad (6)$$

En combinant les équations (5) et (6), nous obtenons que la moyenne arithmétique des vecteurs cepstraux est équivalente au logarithme de la moyenne géométrique des spectres de puissance du signal propre plus un terme représentant le canal téléphonique.

$$\sum_{i=0}^{N-1} C_y^i \Leftrightarrow \log\left(\prod_{i=0}^{N-1} \Gamma_s^i(f)\right) + N\log|H(f)|^2 \qquad (7)$$

Reste à déterminer ce que représente le produit des spectres de puissance du signal propre. Or le spectre de puissance est le produit du spectre à court terme par son conjugé. Si nous croisons adéquatement [6] par couples les spectres et les spectres conjugés dans la multiplication, nous obtenons les spectres croisés de trames du signal non cohérentes. Ces spectres croisés représentent la corrélation ou la corrélation circulaire (selon le calcul des spectres à court terme) des trames. Cette corrélation est généralement connue comme la fonction de cohérence du signal propre. Ainsi la moyenne arithmétique des cepstres du signal représente le logarithme du produit des spectres de cohérence des trames du signal plus N fois le logarithme de la fonction de transfert du canal téléphonique. Si le signal propre est non cohérent, la moyenne des cepstres représente le cepstre du canal.

En se basant sur ce résultat théorique, nous pouvons voire jusqu'où la moyenne des vecteurs cepstraux représente effectivement le canal acoustique. A titre d'exemple, dans les parties de non-parole et comme le silence (ou le bruit de fond) est noncohérent, un horizon limité suffit pour estimer le canal téléphonique. Pour la parole, il faut considérer un horizon beaucoup plus étendu pour estimer le canal téléphonique. Pour les communications à effets de lignes lentement variables, un filtrage passe-haut à temps de réponse élevé suffit pour réduire ces effets dans l'espace de représentation. Ceci justifie les améliorations obtenues dans la littérature [2][4][1].

Les premières expériences sont effectuées sur les corpus de laboratoire: Chiffres et Trégor. Pour chaque appel, un cepstre à long-terme est calculé et soustrait de tous les vecteurs cepstraux pour produire les vecteurs normalisés. Seuls, les 6 premiers coefficients cepstraux sont normalisés. Les expériences montrent que les effets des lignes téléphoniques se limitent aux premiers coefficients cepstraux et sont moins apparents sur les coefficients de rang élevé. Le tableau 2 résume les résultats obtenus. Sur ce tableau sont montrés aussi les intervalles

Tableau 2 : *Résultats de reconnaissance sans et avec déconvolution par soustraction cepstrale.*

Chiffres	taux d'erreur	*réduction du taux d'erreur*
Système de base	0,83%	
Seuil (amélioration significative à 95%)	**0,58%**	*30%*
Après normalisation par réseau	0,44%	*47%*
Trégor		
Système de base	0,64%	
Seuil (amélioration significative à 95%)	**0,52%**	*18%*
Après normalisation par réseau	0,52%	*18%*

de confiance à 95%, prouvant que les améliorations obtenues par soustraction cepstrale apporte une réduction statistiquement significative du taux d'erreur.

Suite aux améliorations obtenues sur différentes bases de données de laboratoire, les techniques de réduction des effets des lignes téléphoniques sont testées sur des données d'exploitation. Le corpus "Les Baladins" est utilisé. La modélisation des mots du vocabulaire se fait par des modèles de mots à 27 états plus deux modèles poubelles qui servent pour le rejet d'entrées incorrectes. Les modèles poubelles sont appris sur le corpus des chiffres isolés et sur une partie du corpus Trégor. Pour tester le rejet, nous utilisons le corpus des nombres à deux chiffres et celui des 500 mots.

Deux techniques pour la réduction des effets convolutifs des lignes téléphoniques sont testées. La soustraction cepstrale déjà décrite et un filtrage RII passe-haut d'ordre 1. Ce filtrage est identique au filtre RASTA

décrit dans la littérature [2][3]. Les cepstres normalisés $\hat{C}_s$ en sortie du filtre sont donnés par l'équation (8).

$$\hat{C}_s^i = C_y^i - C_y^{i-1} + (1-\lambda)\,\hat{C}_s^{i-1} \qquad (8)$$

Le paramètre λ caractérise complètement le filtre RII. Des expériences préliminaires ont permis d'optimiser empiriquement la valeur de λ. Nous avons noté sur nombreux corpus une faible sensibilité à la valeur de ce paramètre dans l'intervalle [0,02; 0,05]. La valeur optimale de λ étant faible, le filtre passe-haut est très sélectif et ne coupe que les très basses fréquences. D'où la confirmation d'un effet de lignes téléphoniques variant lentement dans le temps. Dans la suite nous considérons λ=0.04.

En plus de la normalisation des coefficients cepstraux nous normalisons dans ces expériences le coefficient d'énergie, en échelle logarithmique, par filtrage de sa partie continue et puis l'ajout d'une constante. Ce qui est équivalent à un contrôle automatique de gain.

Les tableaux 3 et 4 résument les résultats obtenus. Les taux d'erreur du système de base sont présentés avec l'intervalle de confiance. En observant ce tableau, nous remarquons que la soustraction cepstrale et le filtrage RII donnent des améliorations statistiquement significatives. Nous remarquons sur cette expérience que le filtrage passe-haut améliore plus que la soustraction cepstrale les performances en reconnaissance. Cette remarque n'est pas vérifiée pour toutes les expériences de reconnaissance que nous avons menées.

VI. CONCLUSIONS

Dans ce papier nous montrons que les lignes téléphoniques agissent sur le signal comme un filtre convolutif linéaire lentement variable au cours d'un appel. Nous arrivons à cette conclusion à travers les mesures effectuées sur des données téléphoniques, et à travers les améliorations obtenues par élimination de la partie continue des trajectoires cepstrales.

Par ailleurs, nous démontrons un théorème qui définit ce que représente la partie continue des vecteurs cepstraux sur un horizon déterminé. Cette partie continue représente l'effet du canal convolutif constant sachant que le signal propre est non cohérent. Ceci justifie les améliorations obtenues par la soustraction cepstrale et celles obtenues par un filtrage RII passe-haut. En effet, tout se joue sur la stabilité des effets du canal contre la grande variabilité (ou la noncohérence) du signal propre. Cette justification théorique est nouvelle par rapport aux travaux rencontrés dans la bibliographie et permet de voir clairement jusqu'où nous pouvons espérer arriver par des techniques de déconvolution avec un seul microphone pour capter le signal de parole.

La déconvolution par transformation des données utilisant un perceptron multicouche apporte une amélioration timide des résultats. Des travaux sont à faire pour optimiser plus le réseau et son utilisation.

Comme principale perspective à ces travaux, nous envisageons d'approfondir le problème dans le cas d'un bruit ambiant additif non négligeable.

Tableau 3 : *Taux d' erreur de reconnaissance en utilisant le filtrage RII ou la soustraction cepstrale.*

Reco. du Vocabulaire.	Test données laboratoire		Test données d'exploitation	
	τ erreur	*réduction*	τ erreur	*réduction*
Système de base	1,14%		3,04%	
Seuil d'amél(95%)	**0,95%**	**17%**	**2,74%**	**10%**
SCep + Nrm. Ener	0,81%	*29%*	2,72%	*11%*
RII + Nrm. Ener	0,86%	*25%*	2,65%	*13%*

Tableau 4 : *Taux de rejet (%mots hors vocabulaire non rejetés) en utilisant le filtrage RII ou la soustraction cepstrale.*

Rejet non vocabulaire	Nbres à 2 chiffres		500 mots	
	τ erreur	*réduction*	τ erreur	*réduction*
Système de base	22,5%		38,6%	
Seuil d'amél(95%)	**21,8%**	**3%**	**37,8%**	**2%**
SCep + Nrm. Ener	21,6%	*4%*	35,4%	*8%*
RII + Nrm. Ener	20,7%	*8%*	28,1%	*27%*

Réference

[1] B.A. Hanson & T.H. Applebaum, "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech," ICASSP, pp. 79-82, 1993.

[2] H. Hermansky, N. Morgan, A. Bayya & P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)," Eurospeech, pp.1367-1370, Genova 1991.

[3] H. Hermansky, N. Morgan, H.G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing," ICASSP, pp. 83-86, 1993.

[4] H.G. Hirsch, P. Meyer & H. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes," Eurospeech, pp. 413-416, Genova 1991.

[5] C. Mokbel, J. Monné & D. Jouvet, "On-Line Adaptation of a Speech Recognizer to Variations in Telephone Line Conditions" Eurospeech, pp. 1247-1250, Berlin 1993.

[6] C. Mokbel, D. Jouvet & J. Monné, " Deconvolution of Telephone Lines Effects for Speech Recognition," soumis à Speech Communication.

APPROCHE COOPÉRATIVE DE L'EXTRACTION DE LA FRÉQUENCE FONDAMENTALE

ALAIN SOQUET[1]

INSTITUT DES LANGUES VIVANTES ET DE PHONÉTIQUE
UNIVERSITÉ LIBRE DE BRUXELLES
50 av F. D. Roosevelt, B-1050 Bruxelles, BELGIQUE

Résumé

Dans ce travail, nous appliquons le principe de l'approche coopérative à l'extraction de la fréquence fondamentale. Nous commencerons par une description de la problématique et des différentes approches actuellement disponibles pour extraire la fréquence fondamentale. Nous montrerons ensuite l'intérêt d'utiliser une approche coopérative, et nous proposerons une méthode simple de combinaison de différents analyseurs. Les performances des différents extracteurs seront ensuite évaluées sur un même corpus.

Mots-clés : fréquence fondamentale, traitement de signal.

INTRODUCTION

Les problèmes liés à l'extraction précise, fiable, et automatique de la période fondamentale de signaux de parole ont été étudiés depuis plus d'un siècle. Cet intérêt résulte de l'importance centrale de cette mesure pour de nombreux systèmes d'analyse et de traitement de la parole.

La mesure automatique de la période fondamentale est une tâche d'une grande complexité. Il faut en chercher les raisons dans les mécanismes mis en œuvre pour produire le signal de parole. Nous allons passer en revue les principales difficultés auxquelles sont confrontées les méthodes d'extraction de la fréquence fondamentale (voir par exemple Hess [4], et Rabiner et al. [11] pour plus de détails).

- *Périodicité* : le système à l'origine des vibrations laryngées est un oscillateur auto-entretenu. L'oscillation observée à la sortie du système n'est pas synchronisée sur une horloge extérieure imposant au système les instants d'ouverture et de fermeture de la glotte. Bien au contraire, ces instants sont le résultat d'une combinaison complexe de phénomènes physiologiques et aérodynamiques qui sont loin de garantir un signal d'excitation parfaitement périodique. Il convient donc d'être prudent en utilisant la relation entre la période fondamentale T_0 et la fréquence fondamentale F_0 :

$$F_0 = \frac{1}{T_0} \tag{1}$$

- *Couplage entre excitation et conduit vocal* : l'onde de pression acoustique est le seul signal facilement accessible. Il est le résultat de la modulation des caractéristiques de la source avec celles du conduit vocal. La séparation des contributions de la source et du conduit pose donc un problème supplémentaire. L'interaction entre la source et le conduit vocal est par exemple importante lors de variations rapides de la forme du conduit, qui s'accompagnent souvent de modifications de la période fondamentale. La combinaison de ces différents facteurs rend la détermination des caractéristiques de la source fort délicate.
- *Segmentation* : une difficulté supplémentaire survient du fait que la recherche de la fréquence fondamentale n'a de sens que si le mécanisme de vibration des cordes vocales est bien en action. Il convient donc de prendre une décision sur le caractère voisé ou non d'une portion de signal de parole; or, la segmentation en zones voisées et non-voisées pose elle-même un problème.

Etant donné les difficultés liées à la mesure de la fondamentale, un grand nombre de travaux lui ont été consacrés et il est possible de se faire une idée de cette diversité en se référant à l'ouvrage de Hess [4].

Les analyseurs fonctionnent généralement en trois étapes successives : (i) un préprocesseur fournit un signal pour lequel l'extraction de la fondamentale est aisée, (ii) un extracteur détermine les valeurs possibles de la fondamentale (par exemple par recherche de maxima, ou de passages par zéro), et (iii) un post-

[1]Bourse de la Banque Nationale de Belgique

traitement sélectionne le candidat retenu comme valeur de la fondamentale.

Les algorithmes peuvent être séparés en trois grandes classes suivant le domaine dans lequel s'effectue l'extraction :

1. *Les analyseurs temporels* cherchent soit à repérer la périodicité directement dans le signal de parole, soit à isoler l'harmonique fondamentale, soit encore à détecter des événements caractéristiques du cycle glottique,
2. *Les analyseurs spectraux* utilisent le fait que le spectre d'amplitude d'un signal périodique sera constitué d'une série d'impulsions à la fréquence fondamentale et en ses harmoniques,
3. *Les analyseurs par corrélation* font l'hypothèse que la forme du signal de parole (ou d'un signal dérivé) varie peu d'une période à l'autre, et qu'il est possible de déterminer la période fondamentale par similarité entre le signal et ce même signal décalé dans le temps. Le décalage donnant la meilleur corrélation indique la période fondamentale.

Dans la section suivante, nous décrivons brièvement les extracteurs utilisés dans ce travail. Ensuite nous introduirons les motivations qui nous ont poussés à utiliser une approche coopérative, et la manière dont les différents extracteurs peuvent être combinés. Les différents analyseurs obtenus seront ensuite évalués sur un même corpus.

MÉTHODES MISES EN ŒUVRE

Nous avons utilisé trois analyseurs différents, chacun appartenant à une classe différente : le critère de variation d'amplitude (une méthode temporelle), le peigne spectral (une méthode spectrale), et la détermination haute résolution du pitch (une méthode basée sur la corrélation). Nous allons très brièvement rappeler les principes sur lesquels reposent ces différentes méthodes.

1. *Critère de variation d'amplitude à court terme* (CVA) (Jospa [5] et [6]) : la méthode est basée sur une observation de l'évolution dans le temps de l'influx d'énergie dans le conduit vocal. Pendant la partie du cycle durant laquelle la glotte est fermée, il n'y a pas d'influx d'énergie et les pertes sont constantes. Lorsque la glotte s'ouvre, les cavités sous-glottiques se connectent au conduit vocal, les pertes augmentent sensiblement, et l'influx d'énergie dans le conduit vocal augmente également pour atteindre son maximum aux environs de l'instant de fermeture de la glotte. Les pertes redeviennent alors prédominantes et le cycle reprend. Ces deux phénomènes impriment à l'enveloppe lissée du signal observé dans une bande de fréquence choisie, une forme en dent de scie. Ces considérations qualitatives qui sont à la base du critère de variation d'amplitude. La figure 1 présente le type de signal obtenu au moyen de la méthode. Le CVA est un signal présentant des maxima aux alentours des instants de fermeture de la glotte. La période fondamentale peut par exemple s'obtenir à partir de l'intervalle entre deux extrema successifs.

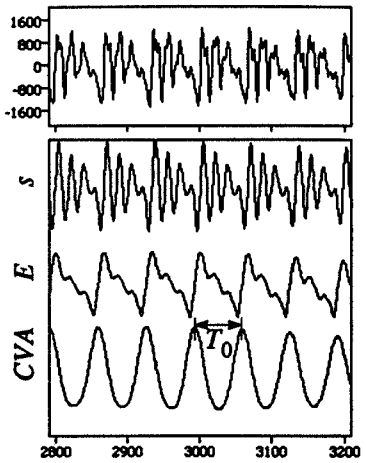


Figure 1 : Application du critère de variation d'amplitude pour une voyelle [a] prononcée par un locuteur masculin. De haut en bas, le signal de parole, le signal filtré dans la bande du premier formant (*S*), l'enveloppe lissée (*E*), et le CVA.

2. *Peigne cepstral* (COMB) (Martin [7] et [8]) : la méthode du peigne consiste à rechercher des valeurs du spectre situées à des fréquences harmoniques et dont la somme est maximale pour un intervalle fréquenciel donné. L'intercorrélation spectre-peigne se réduit au calcul de la somme des composantes spectrales correspondant à une structure harmonique donnée. La structure harmonique donnant le maximum d'intercorrélation est retenue comme

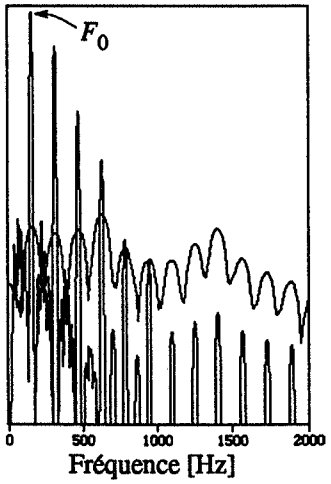


Figure 2 : Application de la méthode du peigne pour une voyelle [a] prononcée par un locuteur masculin. Représentation du spectre d'amplitude et de la fonction d'inter-corrélation d'un peigne à 5 dents.

fréquence fondamentale du signal. La figure 2 présente un exemple d'application de la méthode du peigne.

3. *Détermination haute résolution du pitch* (SRPD) (Medan, Yair, et Chazan [10]) : cette technique est basée sur une modélisation de la similarité entre deux périodes successives d'un signal de parole voisé. Elle fait partie des méthodes fonctionnant par corrélation. Il est supposé que la similarité entre deux périodes successives est forte, et donc que les deux périodes sont des versions modulées en amplitude l'une de l'autre. La figure 3 présente un exemple de fonction d'intercorrélation obtenue par cette méthode. La période fondamentale s'obtient en mesurant la longueur de la fenêtre qui maximise l'intercorrélation.

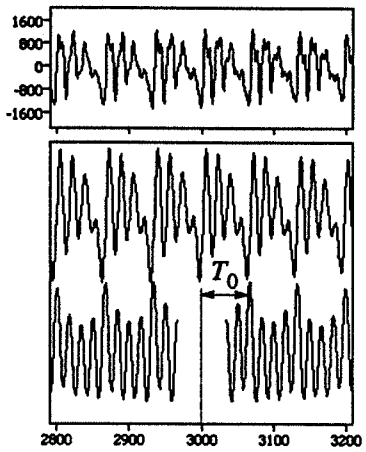


Figure 3 : Application de la méthode de détermination haute résolution du pitch pour une voyelle [a] prononcée par un locuteur masculin. De haut en bas, le signal de parole, le signal filtré passe bas, et l'intercorrélation.

MÉTHODE COOPÉRATIVE

Il existe un très grand nombre de méthodes différentes pour extraire la fréquence fondamentale à partir du signal de parole. Chaque approche présente certains avantages et inconvénients, mais aucune ne permet d'évaluer la fréquence fondamentale avec une fiabilité absolue. Il est de plus relativement peu probable qu'un nouveau principe vienne révolutionner le monde de l'extraction de la fondamentale en résolvant simultanément tous les problèmes.

Ces observations avaient déjà conduit Hess [4] à suggérer de combiner différentes approches pour augmenter les performances globales du système d'extraction. Malgré cette conclusion relativement naturelle, les tentatives de mise en œuvre de ce principe sont restées bien timide.

Gold et Rabiner [2] en 1967 ont proposé un analyseur procédant par mise en parallèle de six extracteurs simples travaillant sur six signaux dérivés des extrema du signal de parole filtré passe-bas. Cet analyseur utilise astucieusement la synergie entre ces extracteurs, mais ne permet pas de bénéficier pleinement du principe de coopération car il n'utilise qu'un seul type d'analyseur.

McGonegal et al. [9] en 1975 ont proposé un outil dont le but était d'obtenir de façon semi-automatique le contour intonatif de phrases. Le programme propose simultanément à l'utilisateur trois représentations synchronisées permettant de localiser interactivement la fréquence fondamentale en un instant donné. Les représentations sont le signal de parole filtré passe-bas, l'autocorrélation de ce même signal, ainsi que le cepstre du signal de parole. Des extracteurs simples proposent à l'utilisateur une valeur de la période fondamentale pour chaque représentation. L'utilisateur interagit alors avec le système pour sélectionner et ajuster la période fondamentale. Dans ce cas, les extracteurs sont bien indépendants, mais l'intégration reste manuelle.

Plus tard, d'autres auteurs ont tenté de sélectionner l'analyseur le mieux adapté à la situation de manière à maximiser les performances globales du système. Citons par exemple Tucker et Bates [12], ainsi que De Mori et al. [1]. Toutefois, les différents extracteurs ne sont pas indépendants : ils travaillent dans le même domaine mais sont adaptés à certains types de signaux ou certaines plages de fréquence.

D'un point de vue général, l'utilisation d'une approche coopérative se justifie par le fait qu'il est peu probable que différentes méthodes de mesure évaluent simultanément de façon erronée une même grandeur. Le principe est donc conceptuellement simple et peut être appliqué de trois manières distinctes :

1. *Sélection* : en fonction de la situation courante, la méthode la plus adaptée est sélectionnée et est utilisée pour donner une estimation de la fondamentale.
2. *Essais-erreurs* : une méthode est appliquée à la situation courante, si l'estimation de la fondamentale faite par cette méthode est jugée erronée, une autre méthode est utilisée; ce processus est répété jusqu'à l'obtention d'une évaluation acceptable.
3. *Combinaison* : différents analyseurs sont appliqués simultanément; les différentes estimations ainsi obtenues sont ensuite combinées.

La méthode travaillant par *sélection* requiert un moyen de choisir à priori l'analyseur le mieux adapté à la situation. Ceci nécessite l'extraction d'indices permettant cette sélection. Cette approche n'est donc pas réaliste car le domaine d'application optimal des différentes méthodes n'est pas connu explicitement, et ce travail de sélection serait d'une complexité équivalente au problème initial.

La méthode travaillant par *essais-erreurs* demande une évaluation de la solution proposée par l'extracteur.

Ceci pose un problème relativement proche de celui posé par l'approche par sélection, à savoir la détection des erreurs des analyseurs. La différence majeure réside dans le fait que, cette fois la détection se fait à posteriori, après application de la méthode. Quoique sensiblement plus réaliste, cette approche n'est que difficilement applicable, car l'évaluation du caractère erroné ou non de l'estimation est un problème qui est loin d'être résolu.

La dernière méthode est basée sur la *combinaison* des estimations fournies par différents analyseurs. Cette approche ne nécessite aucune connaissance sur les domaines d'application des différentes méthodes utilisées. Elle se base directement sur le principe de base de l'approche coopérative énoncé plus haut.

Le principal problème posé par cette approche est justement lié au mécanisme de combinaison. En effet un simple mécanisme de vote de type majoritaire ou autre n'est pas applicable directement, car la variable prend des valeurs continues (à savoir la fréquence fondamentale) et non qualitatives. Nous avons opté pour un mécanisme de combinaison basé sur la corrélation des candidats avec une fenêtre étroite se déplaçant sur le domaine de variation de la fréquence fondamentale. La valeur choisie correspond à la position de la fenêtre pour laquelle la corrélation est maximale.

Soit $C^m(F)$ une fonction prenant une valeur différente de zéro pour les valeurs de F pour lesquelles un candidat est proposé par l'extracteur de la méthode m, soit :

$$C^m(F) = \sum_i C_i^m \delta(F - F_i) \qquad (2)$$

avec C_i^m la pondération du candidat i (variant entre 0 et 1 suivant l'importance attribuée à ce candidat par la méthode) et $\delta(F)$ la distribution de Dirac.

Soit $W(F)$ la fenêtre utilisée pour la corrélation; une fenêtre triangulaire est utilisée et définie comme suit :

$$W(F) = \begin{cases} 1 - F/\Delta F & -\Delta F \le F < 0 \\ 1 + F/\Delta F & 0 \le F < \Delta F \\ 0 & ailleurs \end{cases} \qquad (3)$$

avec ΔF un paramètre égal à la demi-longueur de la fenêtre. Si nous considérons qu'il y a M méthodes différentes, la corrélation peut s'exprimer par :

$$\rho(F) = \sum_{m=1}^{M} \left(\int_{F_0^{min}}^{F_0^{max}} C^m(f) W(F - f) df \right) \qquad (4)$$

La corrélation $\rho(F)$ est définie entre les valeurs minimum F_0^{min} et maximum F_0^{max} autorisées pour la fréquence fondamentale et vaut zéro ailleurs. La figure 4 illustre le calcul de la corrélation $\rho(F)$.

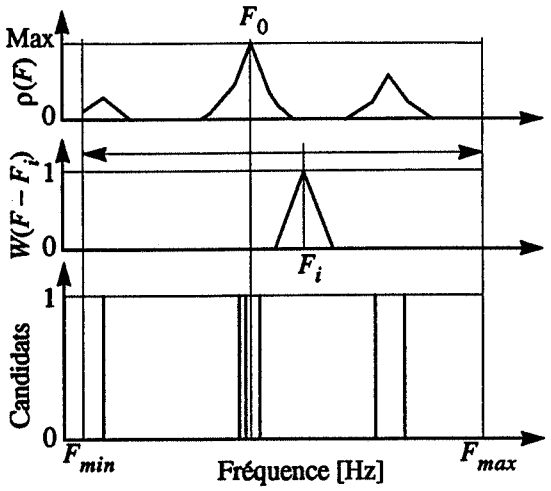


Figure 4 : Corrélation des candidats avec une fenêtre triangulaire.

Nous définissons encore $\eta(F)$ comme le nombre de méthodes différentes présentant un candidat entre $F - \Delta F$ et $F + \Delta F$. Le vote sera considéré comme majoritaire en F si au moins la moitié des méthodes contribuent à la corrélation, soit la condition :

$$\eta(F) > \frac{M}{2} \qquad (5)$$

Le résultat de la combinaison est alors obtenu en cherchant le maximum de la fonction $\rho(F)$ sur le domaine où la condition (5) est vérifiée, soit :

$$F_0 = max(\rho(F)) \quad \eta(F) > \frac{M}{2} \quad F_0^{min} \le F \le F_0^{max} \qquad (6)$$

ÉVALUATION

Les trois analyseurs décrits plus haut (CVA, COMB, et SRPD) ainsi que la méthode coopérative (VOTE) ont été évalués sur un même corpus composé de logatomes VCV, V étant une voyelle parmi [a, œ, i, y, u], et C une occlusive parmi [p, t, k, b, d, g]. Le corpus a été prononcé par trois locuteurs masculins, et segmenté manuellement au préalable de manière à extraire le ou les segments présentant une excitation laryngée. Une mesure de la fréquence fondamentale a été effectuée dans ces segments tous les 100 msec, soit un total de 16541 mesures.

Les catégories d'erreurs répertoriées sont :

- les erreurs résultant du choix de la deuxième harmonique de la fondamentale, soit $F_0 \times 2$,
- les erreurs liées au choix d'une sous-harmonique, soit $F_0/2$,
- une sur-estimation de la fondamentale de plus de 10% soit $\uparrow F_0$,

- une sous-estimation de la fondamentale de plus de 10% soit $\downarrow F_0$,
- les cas pour lesquels la méthode ne propose pas de candidat, nous les noterons $\not F_0$.

Table 1 : Résultat en nombre d'erreur de l'évaluation des méthodes CVA, COMB, SRPD, et VOTE pour un total de 16541 mesures.

Méthode	CVA	COMB	SRPD	VOTE
$F_0 \times 2$	181	23	39	4
$F_0/2$	1	257	432	61
$\uparrow F_0$	30	23	36	12
$\downarrow F_0$	52	14	18	3
$\not F_0$	1913	71	17	0
Total	2184	388	542	80

Les résultats obtenus sont résumés à la table 1 et à la figure 5. Dans le cas du CVA, le taux important d'omission s'explique davantage par la simplicité de l'algorithme utilisé pour passer des instants détectés par la méthode à une évaluation de la fréquence fondamentale que par une déficience de la méthode. Le score très faible de $F_0/2$ s'explique en partie par le fait que la méthode rate rarement une période mais aussi par le taux élevé d'omission. Le COMB est la plus fiable des trois méthodes de base et se comporte très bien dans les différentes catégories d'erreur. Le SRPD, malgré sa grande simplicité, donne des taux d'erreur très acceptables. Seul $F_0/2$ présente un score relativement élevé. Ceci est lié à la pondération de la fonction intercorrélation par l'énergie présente dans la partie utilisée du signal. Ceci implique que la méthode ne défavorise pas l'intercorrélation pour une longueur égale à un multiple de la période.

Pour les trois méthodes, le taux d'erreur non-harmonique ($\uparrow F_0$, et $\downarrow F_0$) est comparable. Ce type d'erreur est particulièrement délicat à corriger par un éventuel post-traitement.

Dans le cas du VOTE, nous pouvons observer une diminution sensible du taux d'erreur. L'erreur totale est diminuée d'un facteur cinq par rapport à la méthode de base la plus performante. L'erreur observée avec le VOTE est toujours inférieure au meilleur score obtenu par les méthodes prises individuellement.

Notons la diminution significative des erreurs non-harmoniques qui sont les plus difficiles à corriger à posteriori. Les erreurs de type $F_0/2$ représentent à elles seules plus de 75% de l'erreur totale. Ce type d'erreur est de surcroît facile à prendre en compte dans un post-traitement simple de type filtre médian.

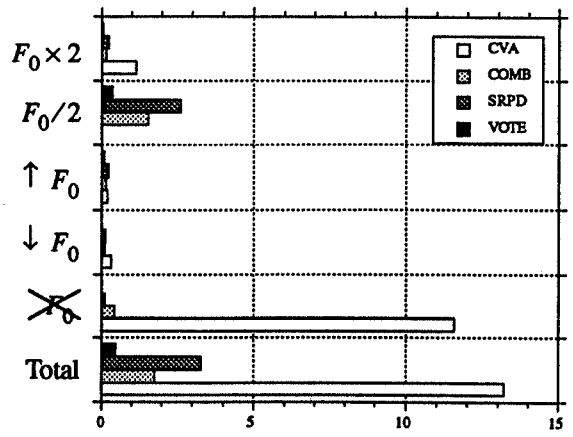


Figure 5 : Résultats en pourcentage de l'évaluation des méthodes CVA, COMB, SRPD, et VOTE.

La figure 6 illustre l'application sans aucun post-traitement des différentes méthodes à une phrase prononcée par un locuteur féminin (voir Hermes [3] pour comparaison). Les observations faites sur le corpus d'évaluation sont clairement visibles. Le COMB apparait clairement comme la plus fiable des 3 méthodes de base. Le VOTE ne commet qu'une erreur grossière (de type $F_0/2$) et propose une courbe mélodique continue et bien lisse.

CONCLUSION

Dans ce travail, nous nous sommes intéressés à l'extraction de la fréquence fondamentale du signal de parole. Après avoir passé en revue différentes approches possibles, nous avons étudié en détail trois analyseurs de type différents : le CVA un analyseur temporel, le COMB un analyseur spectral, et le SRPD un analyseur par corrélation. Nous avons décidé d'aborder l'extraction en utilisant une approche coopérative. Bien que ce principe soit depuis longtemps utilisé avec succès dans d'autres disciplines, une telle approche n'a jusqu'à présent jamais été utilisée systématiquement pour l'extraction de paramètres du signal de parole. Suggérée par Hess [4] dès 1981, la combinaison de plusieurs méthodes basées sur des principes différents n'avait pas encore été tentée. Nous avons donc présenté une manière simple et efficace de combiner différentes méthodes, et nous l'avons évaluée sur un corpus de logatomes prononcés par trois locuteurs. Les résultats observés sont très nettement en faveur de l'approche coopérative et laissent entrevoir une utilisation prometteuse dans le cas de l'extraction automatique d'autres indices acoustiques du signal de parole.

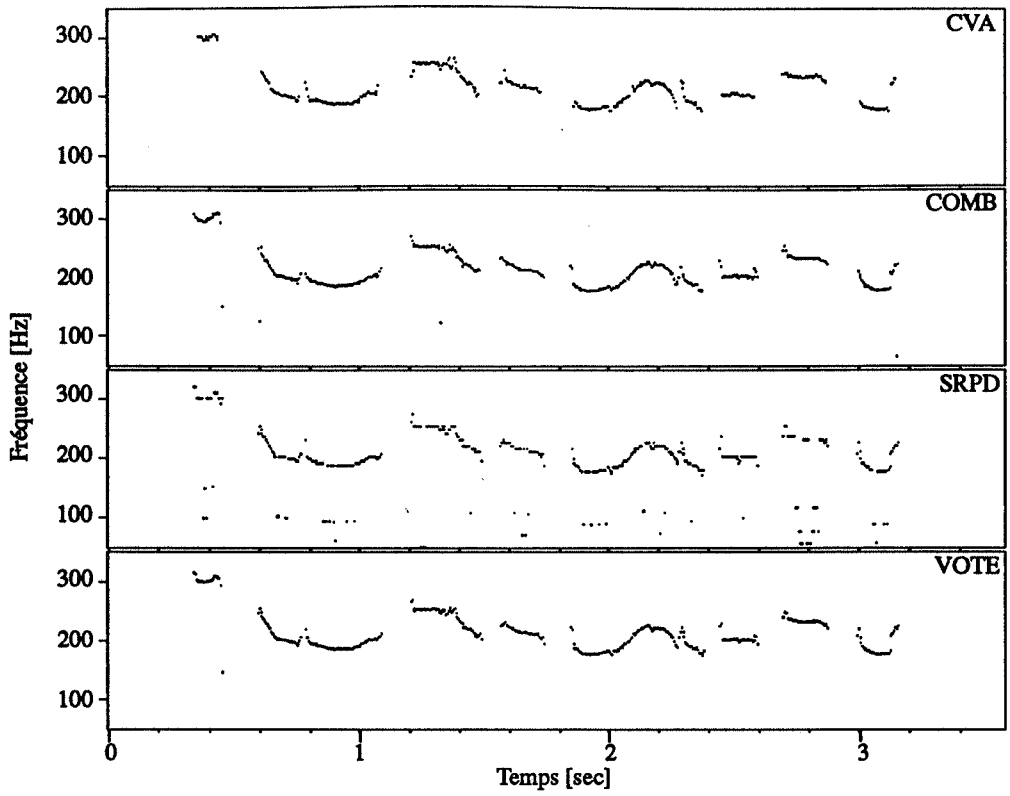


Figure 6 : Application des méthodes CVA, COMB, SRPD, et VOTE sans post-traitement à la phrase "Fred can go, Susan can't go, and Linda is uncertain." prononcée par un locuteur féminin.

RÉFÉRENCES

[1] R. De Mori, P. Laface, V. A. Makhonine, and M. Mezzalama, "A syntactic procedure for the recognition of glottal pulses in continuous speech," Pattern Recognition, vol. 9, pages 181-189, 1977.

[2] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. Am., vol. 46, pages 442-448, 1969.

[3] D. J. Hermes, "Pitch analysis," in Visual representations of speech signals: M. Cooke, S. Beet and M. Crawford (eds.), John Wiley & Sons Ltd, pages 3-25, 1993.

[4] W. J. Hess, "Pitch determination of speech signals - Algorithms and devices," Springer-Verlag Berlin, 1983.

[5] P. Jospa, "Critères de variation d'amplitude à court terme," Rapport d'activités de l'Institut de Phonétique de l'Université Libre de Bruxelles, vol. 17, pages 89-108, 1982.

[6] P. Jospa, "Détection synchrone du pitch par le critère de variation d'amplitude (CVA)," Actes des 13èmes Journées d'études sur la parole, Bruxelles, pages 161-162, 1984.

[7] P. Martin, "Une nouvelle méthode de mesure de la fréquence fondamentale par intercorrélation avec une fonction peigne," Rapport d'activités de l'Institut de Phonétique de l'Université Libre de Bruxelles, vol. 15, pages 117-128, 1981.

[8] P. Martin, "Une méthode de calcul rapide du peigne spectral pour la mesure de la fréquence fondamentale," Travaux de l'Institut de Phonétique d'Aix, vol. 10, pages 361-369, 1985.

[9] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Semi-Automatic Pitch Detector (SAPD)," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, pages 570-574, 1975.

[10] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," IEEE Transaction on Signal Processing, vol. 39, n° 1, pages 40-48, 1991.

[11] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, n°5, 1976.

[12] W. H. Tucker and R. H. T. Bates, "A pitch estimation algorithm for speech and music," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, pages 507-604, 1978.

Extraction des formants en synchronie avec l'excitation à l'aide d'un modèle auto-régressif composé

Z. Azami* et J. Schoentgen†

Institut des Langues Vivantes et de Phonétique, CP110, Université Libre de Bruxelles, Av. F. D. Roosevelt, 50, B-1050 Bruxelles, Belgique

Résumé

Une nouvelle méthode d'extraction des formants, en synchronie avec l'excitation, est présentée. Elle permet d'obtenir une estimation des fréquences formantiques sur la partie du signal qui correspond à la phase fermée du cycle glottique. Conventionellement, les positions et durées des phases fermées sont déterminées au moyen d'un signal auxiliaire tel que le laryngogramme, par exemple, avant qu'un modèle auto-régressif y soit ajusté. Un des problèmes que pose l'approche conventionnel est que les instants d'ouverture et de fermeture de la glotte doivent être connus. La méthode présentée ici ne requiert pas que les positions et durées des phases fermées soient connues a priori. Elle utilise un modèle auto-régressif composé. Celui-ci est constitué de deux sous-modèles linéaires auto-régressifs qui sont ajustés séparément au signal: le premier lors de la phase ouverte et le deuxième lors de la phase fermée. Les positions et durées de celles-ci sont déterminées en minimisant l'erreur de prédiction totale. Des résultats obtenus sur des phrases naturelles sont présentés. Ceux-ci montrent que la segmentation en phases fermées/ouvertes est robuste et que les fréquences et largeurs de bande obtenues sont typiquement celles qui correspondent aux phases fermées du cycle glottique.

Mots-clefs : extraction des formants en synchronie avec l'excitation, modèle auto-régressif composé.

Introduction

L'extraction des formants, en synchronie avec l'excitation, est l'estimation des fréquences de résonances du conduit vocal pendant la phase fermée de la glotte. Celle-ci suit immédiatement l'excitation principale du conduit.

Les fréquences formantiques ainsi obtenues sont réputées fournir la meilleure estimation des fréquences des pics de la fonction de transfert du conduit vocal. En effet, durant la phase fermée, le signal de source est faible ou zéro. Le signal de parole est alors l'evolution libre de la réponse du conduit vocal. Les hypothèses mathématiques qui garantissent l'adéquation du modèle linéaire auto-régressif sont alors le mieux vérifiées.

La définition d'un modèle linéaire auto-régressif est la suivante.

$$y(n) = a_0 + a_1 y(n-1) + a_2 y(n-2) + \cdots + a_N y(n-N) + z(n). \quad (1)$$

$y(n)$ est la réponse du conduit, a_i les coefficients du modèle, $z(n)$ le résidu et N l'ordre du modèle.

Les valeurs des coefficients a_i sont estimées en minimisant l'erreur de prédiction (2).

$$\sum_n \epsilon^2(n) = \sum_n \{y(n) - a_0 - a_1 y(n-1) - a_2 y(n-2) - \cdots - a_N y(n-N)\}^2 . \quad (2)$$

Dans le plan z, le modèle (1) est transformé en un polynôme dont les arguments des racines complexes sont les estimations des fréquences des formants du signal $y(n)$ [3].

Un des problèmes majeurs que pose l'extraction des formants en synchronie avec l'excitation est que les instants d'ouverture et de fermeture de la glotte doivent être connus avant que le modèle (1) ne puisse être ajusté. Le problème de la détection automatique de ces instants, à partir du signal de parole ou d'un signal auxiliaire, n'a pas encore trouvé de solution satisfaisante à l'heure actuelle [4].

Un signal auxiliaire souvent utilisé est le laryngogramme [2]. Celui-ci évolue de façon inversément proportionnelle à la surface de contact des cordes vocales. Le pic négatif de la dérivée du laryngogramme indique alors le début de la fermeture de la glotte. Cette approche pose plusieurs problèmes. D'abord, l'impédance glottique évolue graduellement lors de l'ouverture de la glotte. Par conséquent, l'ouverture ne laisse pas de marque particulière dans le laryngogramme. Ensuite, signal de parole et laryngogramme ne sont pas alignés étant donné que le son se propage

* Boursier U.L.B.
† Fond National de la Recherche Scientifique, Belgique

avec une vitesse finie de la glotte au microphone. C'est pourquoi, les deux signaux doivent être resynchronisés avant que le premier ne puisse être utilisé pour segmenter le second. Cependant, la possibilité de resynchroniser présuppose l'existence d'indices dans le signal de parole, à l'aide desquels les deux signaux puissent être alignés. Or, la raison pour laquelle on a recours au laryngogramme est justement que ces indices sont difficiles à trouver ou inexistants. Aussi, même si la synchronisation était possible, elle devrait être un processus dynamique, étant donné que la longueur du conduit vocal évolue continuellement et que la vitesse du son est différente à l'intérieur et à l'extérieur. Enfin, le laryngogramme ne peut pas être obtenu pour tout locuteur : une impédance du cou trop élevée empêche de capter le laryngogramme avec un rapport signal/bruit suffisant.

Nous proposons ici de contourner le problème de la détermination a priori des différentes phases du cycle par une méthode qui consiste à ajuster au signal de parole deux modèles linéaires auto-régressifs : le premier représente la portion du signal qui correspond à la phase ouverte, le deuxième la portion qui correspond à la phase fermée. La formulation du modèle repose sur le raisonnement suivant. Un segment de parole voisée peut être localement représenté par la convolution d'un modèle linéaire (1) de la réponse impulsionnelle du conduit vocal avec un modèle (non linéaire) du signal glottique. Un modèle du signal glottique bien connu est le modèle de Liljencrants-Fant [1]. Celui-ci représente la dérivée de l'impulsion glottique à l'aide de deux courbes. La première courbe est une sinusoïde multipliée par une exponentielle. Elle est une solution d'une équation aux différences linéaire du deuxième ordre (3). La deuxième est une exponentielle décroissante. Elle est la solution d'une équation aux différences linéaire du premier ordre (4). Le basculement de l'une vers l'autre se fait au début et à la fin de la phase de retour du signal glottique. Le début de la phase de retour correspond à l'instant où la dérivée du signal glottique est la plus négative.

$$s(n) = a_0 + a_1 s(n-1) + a_2 s(n-2) \quad (3)$$
$$0 \leq n\Delta t \leq T_1,$$

$$s(n) = b_0 + b_1 s(n-1) \quad (4)$$
$$T_1 < n\Delta t < T.$$

$s(n)$ est le signal de source dérivé, n l'indice temporel, Δt le pas d'échantillonnage, T la période et T_1 la durée de la phase ouverte.

La convolution de ce modèle avec un modèle de la réponse impulsionnelle du conduit vocal (1) fournit un modèle du signal de parole $y(n)$.

$$y(n) = c_0 + c_1 y(n-1) + \cdots + c_N y(n-N) + z_c(n)$$
$$0 \leq n\Delta t \leq T_1, \quad (5)$$

$$y(n) = d_0 + d_1 y(n-1) + \cdots + d_M y(n-M) + z_d(n)$$
$$T_1 < n\Delta t < T. \quad (6)$$

$y(n)$ est le signal de parole, c_i et d_i sont les coefficients des sous-modèles, N et M les ordres, n l'indice temporel, $z_c(n)$ et $z_d(n)$ les résidus, Δt le pas d'échantillonnage, T_1 et T respectivement la durée de la phase ouverte et de la période fondamentale.

Les relations (5) et (6) constituent un modèle qui représente séparément les portions du signal de parole correspondant aux phases ouverte et de retour. Bien que ce modèle soit plus complet que le modèle (1), le problème de la détection des instants où les modèles basculent de l'un vers l'autre reste entier. Néanmoins, ce modèle constitue un pas dans la bonne direction. En effet, son ajustement devrait aboutir à la minimisation de l'erreur de prédiction totale sur la durée d'un cycle glottique. Cela veut dire que les durées T_1 et T ne sont pas arbitraires. Au contraire, leurs valeurs sont évaluées, en même temps que celles des coefficients c_i et d_i, lors du calcul de l'erreur totale de modélisation qui doit être minimisée.

Cependant, les durées T_1 et T ne peuvent pas être déterminées directement de façon automatique. La raison en est la suivante. Le nombre de périodes ainsi que le type du premier sous-modèle au sein de la fenêtre d'analyse sont inconnus lorsque celle-ci est positionnée de façon non synchrone avec le signal. C'est pourquoi, il n'est pas possible de segmenter le signal automatiquement à l'aide d'une méthode d'optimisation courante.

Nous proposons ci-après une méthode indirecte qui ajuste automatiquement le modèle composé (5) et (6) du signal de parole compris dans une fenêtre d'analyse positionnée de façon non synchrone. En effet, plutôt que de découper "verticalement" le signal en segments successifs, la méthode le découpe "horizontalement", à l'aide d'un seuil, en deux tranches. En effet, un signal périodique qui possède un pic unique par cycle peut être séparé horizontalement en deux tranches par un seuil dont la valeur est comprise entre le maximum et le minimum. Le premier sous-modèle (5) est alors ajusté à la portion "supérieure" et le deuxième sous-modèle (6) à la portion "inférieure".

A ce stade, deux difficultés subsistent. Premièrement, un simple seuil découpe un signal en tranches dont les débuts et fins coïncident nécessairement avec les instersections du seuil et du signal. Deuxièmement, le signal de parole possède plusieurs maxima

locaux par période et n'est donc pas découpé par un seuil en deux mais en plusieurs tranches. La première difficulté est surmontée en retardant l'effet du croisement du seuil. Les tranches de signal découpées peuvent alors être positionnées partout par rapport aux instants de croisement. Le retard d gère aussi le décalage éventuel du signal de parole et du signal auxiliaire. La deuxième difficulté est résolue en basant la segmentation du signal de parole sur un signal auxiliaire qui possède un seul pic par période (Fig. 1). Le signal auxiliaire doit avoir la même période que le signal de parole, mais n'est pas obligatoirement aligné avec celui-ci.

On obtient ainsi le modèle auto-régressif auto-excité à seuil du signal de parole.

Modèle auto-régressif auto-excité à seuil

On considère le modèle auto-régressif composé suivant.

$$y(n) = c_0 + c_1 y(n-1) + ... + c_N y(n-N) + z_c(n), \quad w(n-d) < r \tag{7}$$

$$y(n) = d_0 + d_1 y(n-1) + ... + d_M y(n-M) + z_d(n). \quad w(n-d) \geq r \tag{8}$$

$y(n)$ est le signal de parole, $w(n)$ un signal auxiliaire, c_i et d_i sont les coefficients des sous-modèles (7) et (8), N et M les ordres, n l'indice temporel, $z_c(n)$ et $z_d(n)$ les résidus, r un seuil et d un délai.

Le signal auxiliaire $w(n)$ a la même période que le signal $y(n)$. Des exemples de signaux auxiliaires sont le laryngogramme et le signal de parole lissé. Les différences entre l'usage conventionnel du signal auxiliaire et l'usage qui en est fait ici sont les suivantes. Premièrement, la recherche de la meilleure segmentation du signal ne se fait pas à l'aide de marques dans le signal auxiliaire mais sur base de l'erreur totale de modélisation du signal de parole. Deuxièmement, signal auxiliaire et signal de parole ne doivent pas être alignés. Troisièmement, les instants de basculement d'un sous-modèle vers l'autre sont déduits non pas d'événements critiques mais de seuils critiques. Quatrièmement, le délai d et le nombre de seuils ne dépendent ni de la position relative de la fenêtre d'analyse par rapport au signal, ni de sa longueur. Par conséquent, la recherche de la meilleure hauteur de seuil r et du meilleur délai d peut être automatisée.

Dans la littérature statistique le modèle (7) & (8) est connu sous le nom de modèle auto-régressif auto-excité à seuil (ARAES) [6].

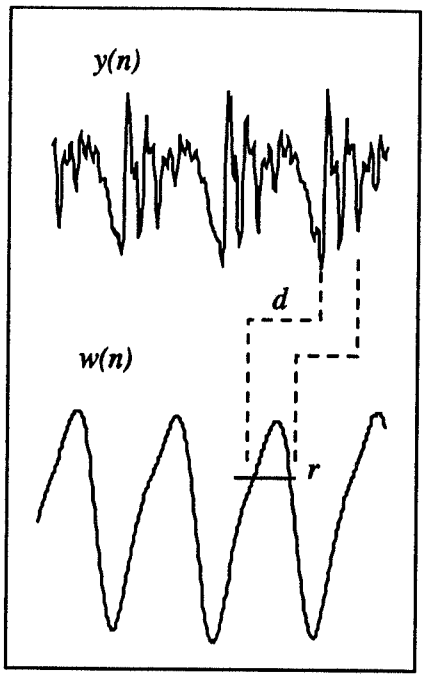


FIG. 1 - *Signal de parole y(n) et signal auxiliaire w(n): action du seuil r et du délai d qui isolent la phase fermée du signal y(n).*

Méthode

Le signal était considéré comme stationnaire au sein de la fenêtre d'analyse. Cette fenêtre était positionnée de façon asynchrone et elle devait avoir une longueur supérieure à la durée d'une période du signal. En pratique, sa durée était de 20 à 30 msec.

Le signal auxiliaire $w(n)$ était construit en lissant le signal de parole compris dans la fenêtre. La segmentation d'un cycle du signal de parole en deux tranches, la première correspondant à la phase ouverte et la deuxième à la phase fermée, était alors effectuée au moyen d'un seuil et d'un délai. Les meilleures seuils et délais étaient déterminés par un optimiseur qui minimisait l'erreur de prédiction totale normalisée des sous-modèles (7) & (8). Lors de cette étape, les valeurs des coefficients c_i et d_i furent estimées en résolvant les systèmes d'équations linéaires par décomposition en valeurs singulières (Singular Value Decomposition, SVD). Les ordres des sous-modèles (7) & (8) étaient respectivement 8 et 10 pour des signaux échantillonnés à 8 kHz. Une fois le minimum de l'erreur atteint, les fréquences formantiques étaient estimées en réajustant le modèle (8) à l'aide d'une méthode de prédiction linéaire covariante multi-intervalles [7]. Les fréquences étaient obtenues en calculant les arguments des racines complexes du polynôme $\sum_{i=0}^{M} d_i z^i$.

L'algorithme était le suivant.

(i) Positionnement asynchrone d'une fenêtre d'analyse de 25 ms.

(ii) Calcul du signal auxiliaire $w(n)$ par lissage du signal de parole.

(iii) Initialisation des valeurs du seuil r et du délai d.

(iv) Segmentation du signal de parole selon les équations (7) & (8).

(v) Ajustement des sous-modèles (7) & (8) sur les segments obtenus.

(vi) Calcul de l'erreur totale de prédiction normalisée.

(vii) Calcul par un optimiseur, sur base de cette erreur, d'une nouvelle valeur pour le seuil r et le délai d. Lorsque l'erreur ne diminuait plus, continuation à (*viii*), sinon reprise en (*iv*).

(viii) Réestimation des valeurs des coefficients d_i par une méthode de prédiction linéaire covariante multi-intervalles.

(ix) Calcul des valeurs des fréquences formantiques à partir des racines complexes du polynôme $\sum_{i=0}^{M} d_i z^i$.

Résultats

La figure 2 montre la phrase "(She h)ad your dark suit in greasy wash water all yea(r)". Les graphiques juxtaposés sont ceux du signal acoustique et des valeurs des fréquences formantiques estimées sur les phases fermées. Les formants dont la largeur de bande estimée était > 500 Hz ont été éliminés. Aucun autre post-traitement n'a été effectué. On voit que les trajectoires formantiques des portions voisées sont plausibles et que les omissions, insertions et valeurs aberrantes sont peu nombreuses. La Figure 3 montre le signal acoustique et les trajectoires formantiques de la phrase "(A la h)âte, le métayer ensilait ses réco(ltes)". Les trajectoires ont été obtenues de la même manière et les résultats sont d'une qualité similaire.

La Figure 4 montre dans l'ordre, le signal de parole, les tranches du signal qui ont été associées par le modèle ARAES avec les phases fermées du cycle glottique, la fréquence fondamentale et les valeurs brutes des fréquences des formants. La portion de signal représentée est [wɔ:(tər)] qui figure dans la phrase montrée à la figure 2. On voit que le modèle ARAES a isolé les segments du signal qui suivent immédiatement l'excitation qui a lieu au moment de la fermeture de la glotte. Une exception possible est le cycle correspondant à t=2.1 s.

Discussion et conclusion

i) Les résultats montrent que les estimations des fréquences formantiques obtenues à l'aide du modèle ARAES sont d'une qualité qui est au moins égale aux résultats obtenus par d'autres méthodes [2].

ii) Actuellement, deux méthodes distinctes de résolution de systèmes linéaires sont utilisées : la décomposition en valeurs singulières et la méthode de covariance. La première gère mieux les exceptions (c.-à-d. déterminant proches de zéro, nombre de données trop réduit etc.) et la deuxième procure des estimations des fréquences formantiques plus fiables.

iii) La méthode d'extraction en synchronie basée sur le modèle ARAES est totalement automatique et opère au sein d'une fenêtre d'analyse positionnée de façon asynchrone.

iv) Les applications envisagées concernent le filtrage glottique inverse en synchronie avec l'excitation du conduit vocal.

Références

[1] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of the glottal flow. *STL-QPSR*, 4:1–13, 1985.

[2] A. K. Krishnamurthy and D. G. Childers. Two-channel speech analysis. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-34(4):730–743, 1986.

[3] S. McCandless. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP-22(2):135–141, 1974.

[4] S. Parthasarathy. Excitation-synchronous modeling of voiced speech. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP-35(9):1241–1249, 1987.

[5] J. Schoentgen. Modelling the glottal pulse with a self-excited threshold auto-regressive model. *Proceedings Eurospeech'93*, Berlin:107–110, 1993.

[6] H. Tong and K. S. Lim. Threshold autoregression, limit cycles and cyclical data. *J. Roy. Stastist. Soc.*, B(42):245–292, 1980.

[7] S. Wood. The precision of formant frequency measurement from spectrograms and by linear prediction. *STL-QPSR*, 1:91–93, 1989.

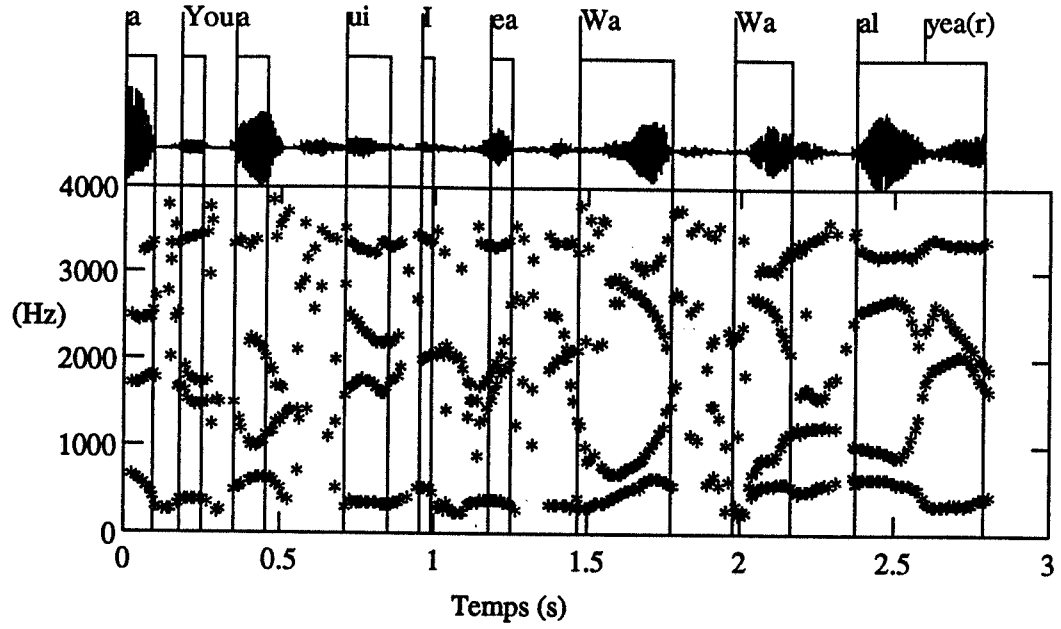


FIG. 2 - *Signal de parole et trajectoires formantiques estimées sur les phases fermées: "(She h)ad your dark suit in greasy wash water all yea(r)"*

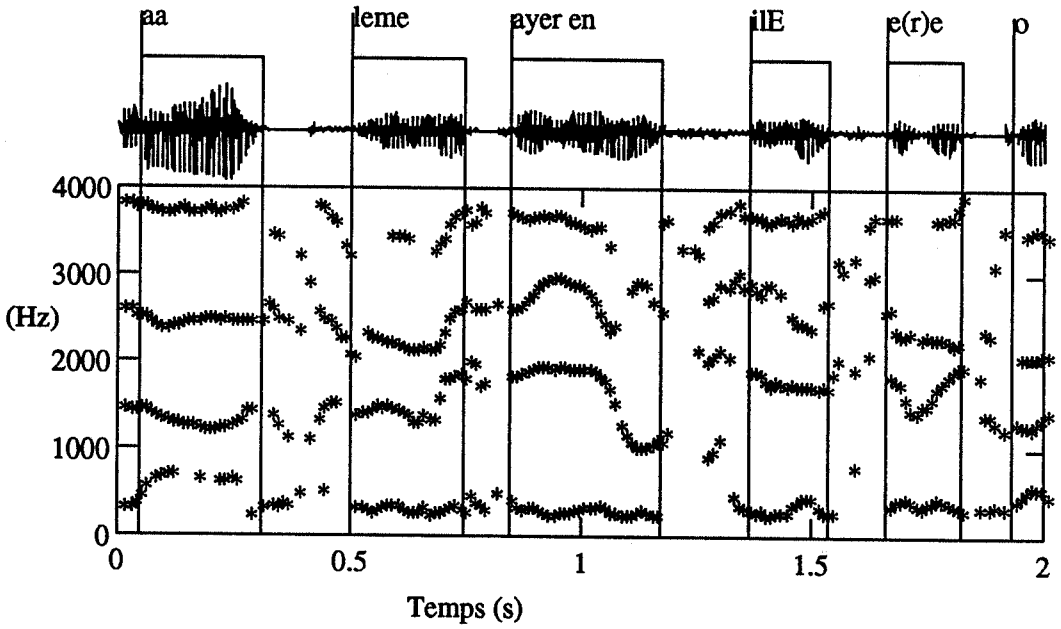


FIG. 3 - *Signal de parole et trajectoires formantiques estimées sur les phases fermées: "(A la h)âte, le métayer ensilait ses réco(ltes)"*

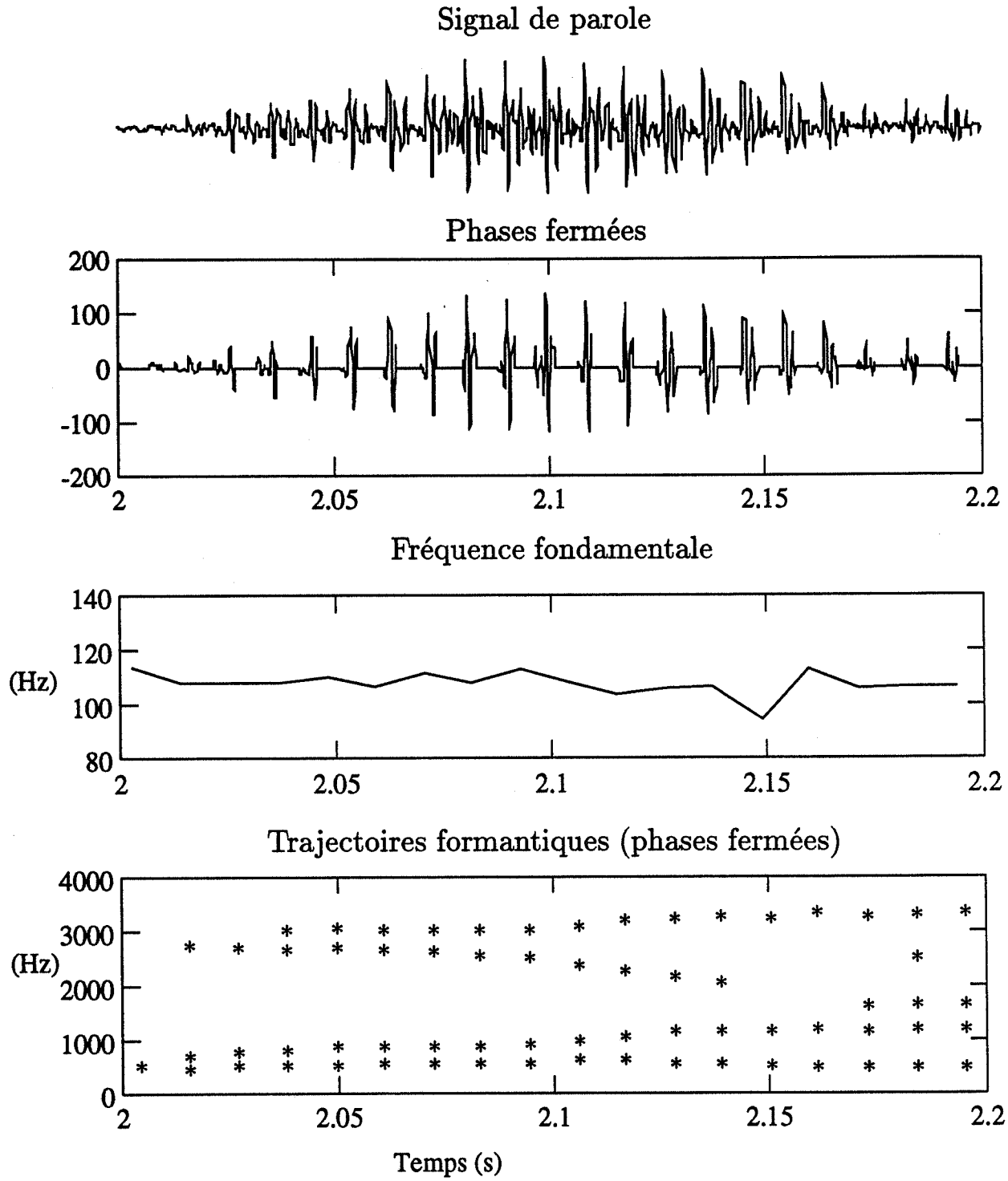


FIG. 4 - *Signal : [wɔ:(tər)]*

CODAGE DE PAROLE CELP À EXCITATION VECTORIELLE CODÉE PAR TREILLIS

Andrei POPESCU [1] Claude LAMBLIN [2] Nicolas MOREAU [1]

[1] TELECOM Paris , 46 rue Barrault, 75013 PARIS FRANCE
[2] CNET LAA/TSS/CMC , route de Trégastel, 22301 LANNION FRANCE

Résumé

Nous utilisons la quantification vectorielle codée par treillis (TCVQ) pour l'excitation d'un codeur CELP dans le but d'obtenir une meilleure qualité du signal quantifié tout en maintenant une complexité de calcul raisonnable. La méthode que nous proposons ici peut être utilisée pour remplacer la quantification vectorielle (VQ) par la TCVQ pour l'excitation de tout codeur CELP standard. Nous proposons un algorithme d'optimisation pour peupler le treillis et nous évaluons la complexité et la performance de la méthode.

Mots-clés:*codage de parole, quantification codée par treillis.*

1. INTRODUCTION

Le codage par treillis et par arbre ont été largement utilisés dans des applications comme la compression de données et la correction d'erreurs. Plusieurs méthodes de codage par treillis ou par arbre ont été proposées pour le signal de parole. Certaines avaient pour but une réduction de complexité [12] [13], d'autres une meilleure qualité, comme la PTWC de Ayanoğlu et Gray [8] ou la FSVQ de Foster, Gray et Dunham [9]. La quantification vectorielle prédictive codée par arbre ou treillis a au moins deux avantages par rapport à la quantification vectorielle prédictive utilisant la même longueur de vecteur. Premièrement, elle rend possible une minimisation de la distorsion sur une période de temps dépassant la longueur d'un vecteur et par conséquent l'effet qu'aura un vecteur d'excitation sur le signal codé qui suit peut être pris en compte dans la détermination de l'excitation; cette propriété est utilisée dans le codeur à décision retardée de Mano et Moriya [10] et dans le codeur à faible retard utilisant une quantification vectorielle par arbre de Shoham [11]. Deuxièmement, avec un codage par treillis ou par arbre, on peut utiliser un dictionnaire contenant plus de vecteurs que celui dont on dispose pour la quantification vectorielle au même débit; pour le codage des sources sans mémoire on démontre que ce dictionnaire plus large permet d'effectuer une quantification plus précise [2].

Exploitant la dualité qui existe entre la modulation pour des canaux perturbés par un bruit gaussien et la quantification des sources sans mémoire, Marcellin et Fischer ont introduit la "Quantification Codée par Treillis" (TCQ) [3], utilisant la théorie de la "Modulation Codée par Treillis" d' Ungerboeck [6]. Depuis, la TCQ a évolué. Il a été montré comment associer la prédiction linéaire avec la TCQ. Dans le cas général, les branches du treillis peuvent être étiquetées avec des vecteurs, donnant naissance à la quantification vectorielle codée par treillis (TCVQ)[2]. La TCQ et la TCVQ ont été utilisées pour la compression du signal de parole dans, respectivement, [4] et [1]. La TCVQ réunit les deux bonnes propriétés des codeurs par treillis/arbre que nous avons citées auparavant.

Nous utilisons la TCVQ pour l'excitation d'un codeur CELP dans le but d'obtenir une meilleure qualité du signal quantifié tout en maintenant une complexité de calcul raisonnable. Nous appellerons le codeur CELP à excitation quantifiée par TCVQ le TCVQ CELP. Les principales différences entre notre approche et celle de [1] sont que nous utilisons la TCVQ dans une structure CELP (même si la structure "noise-feedback" de [1] est similaire au CELP) et que nous calculons le gain d'excitation en boucle fermée. Nous proposons également des algorithmes pour peupler le treillis et une analyse de la complexité et performance de la TCVQ relativement à la complexité et performance de la VQ pour l'excitation du même codeur CELP.

2. LA QUANTIFICATION CODÉE PAR TREILLIS DE L'EXCITATION D'UN CODEUR CELP

2.1. La Structure du Treillis

Tous les treillis pour modulation d'amplitude d'Ungerboeck peuvent être utilisés dans notre méthode. Les branches du treillis sont étiquetées avec des sous-ensembles d'un dictionnaire contenant des vecteurs d'excitation. On peut aussi considérer que chaque branche du treillis est associée à un seul vecteur d'excitation: dans ce cas il y a plusieurs branches parallèles (une pour chaque vecteur du sous-ensemble) entre deux nœuds successifs du treillis. Pourtant, pour la simplicité et généralité de la représentation, nous avons choisi d'étiqueter les branches avec des sous-ensembles du dictionnaire;

les nœuds du treillis sont associés à des états du codeur. Tous les treillis pour modulation d' amplitude d'Ungerboeck correspondent à des codeurs convolutionnels de taux 1/2. Il y a deux branches quittant chaque nœud et également deux branches arrivant dans chaque nœud de ces treillis. Le treillis pour modulation d'amplitude à 8 états est présenté figure 1.

Les m bits disponibles pour la quantification de chaque vecteur d'excitation sont utilisés de la manière suivante. Etant donné un état de départ, $\tilde{m} = 1$ bit, appelé bit de transition, est utilisé en tant que bit d'entrée du codeur convolutionnel, déterminant une transition d'état représentée par une branche dans le treillis. Parallèlement, le codeur convolutionnel produit une paire de bits de sortie, qui identifie le sous-ensemble du dictionnaire étiquetant cette branche, donc il y a quatre sous-ensembles.. Les $m - \tilde{m} = m - 1$ bits restants adressent un vecteur dans le sous-ensemble donné. Par conséquent, le dictionnaire du treillis (l'union des sous-ensembles étiquetant les branches du treillis) contient $4 \cdot 2^{m-1} = 2^{m+1}$ vecteurs, deux fois le nombre de vecteurs contenus dans le dictionnaire de quantification vectorielle pour le même débit d'information.

Pour chacun des codeurs convolutionnels d'Ungerboeck, il existe un codeur équivalent sans contre-réaction; l'utilisation de ce dernier a pour effet qu'une erreur sur un bit de transition ne conduit qu'à un nombre limité de vecteurs erronés au décodeur.

Les branches du treillis peuvent être étiquetées avec plus de 4 sous-ensembles (8, 16, ...)[2], rendant ainsi le dictionnaire du treillis encore plus large; pourtant, dans nos simulations du codeur TCVQ CELP, les treillis étiquetés à plus de 4 sous-ensembles ne se sont pas montrés significativement meilleurs que les treillis étiquetés à 4 sous-ensembles.

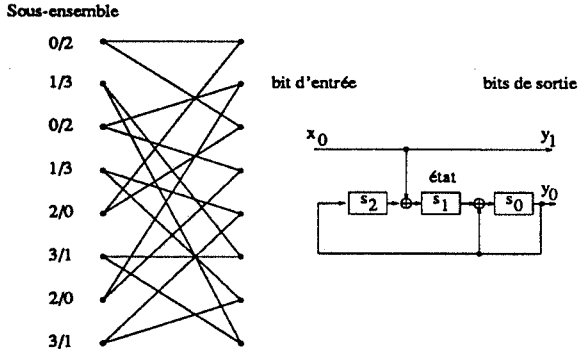


Figure 1. Le treillis à 8 états pour modulation d'amplitude d'Ungerboeck. Le codeur convolutionnel associé.

2.2. La Procédure de Codage TCVQ CELP

L'excitation du codeur consiste en une séquence de vecteurs d'excitation, correspondant à une suite de branches - un chemin - dans le treillis. Le codeur utilise l'algorithme de Viterbi [7] pour trouver une bonne[1] séquence de vecteurs d'excitation parmi les séquences permises par la structure du treillis, qu'on vient de décrire. A un moment donné, avant la quantification du vecteur d'entrée suivant, chaque chemin finit dans un nœud du treillis (état). Ces différents chemins correspondent à différentes séquences d'excitation, et implicitement à différents signaux quantifiés. Une distorsion est associée à chaque chemin, égale à la distance entre le signal original et le signal codé jusqu'à l'instant courant, pondérée pour prendre en compte les caractéristiques psychoacoustiques de l'oreille. Pour la quantification du nouveau vecteur d'entrée, les prédictions à court terme et à long terme sont calculées pour chaque chemin de la même manière que dans le codeur CELP standard (voir [5]) ayant le même signal d'excitation dans le passé. Explicitement, à chaque chemin i correspond un vecteur "cible" t:

$$t = p - p^i_{st} - p^i_{lt} \qquad (1)$$

où p est le vecteur d'entrée filtré par le filtre de pondération perceptuelle (le "filtre perceptuel"), p^i_{st} est le vecteur de prédiction à court terme pour le chemin i filtré par le filtre perceptuel et p^i_{lt} est le vecteur de prédiction à long terme pour le chemin i filtré par le filtre perceptuel. Puis, une recherche en boucle fermée est effectuée pour trouver le meilleur vecteur d'excitation et le gain associé dans chacun des deux sous-ensembles étiquetant les branches partant du nœud final du chemin donné. Donc, si c est un vecteur d'excitation et y est le résultat du filtrage de c à travers le filtre de synthèse à mémoire nulle suivi par le filtre perceptuel à mémoire nulle, dans chaque sous-ensemble on choisit le vecteur d'excitation qui minimise:

$$|t - g_{opt} \cdot y| \qquad (2)$$

Le gain d'excitation optimal correspondant est:

$$g_{opt} = \frac{< t, y >}{|y|^2} \qquad (3)$$

Chaque chemin de départ est prolongé avec ces nouvelles branches, donnant naissance à deux nouveaux chemins conduisant à deux nouveaux nœuds finaux. La distorsion associée à un nouveau chemin est égale à celle du chemin de départ plus la distorsion engendrée par le nouveau vecteur codé. Entre deux chemins conduisant au même nœud final, seulement celui qui donne lieu à la moindre distorsion est gardé pour la suite, l'autre est éliminé. Il en résulte qu'il y a au maximum un chemin conduisant à chaque nœud final, et que le nombre de chemins pris en compte pour la quantification du vecteur d'entrée suivant ne dépasse jamais le nombre d'états du treillis. La procédure énoncée ci-dessus est répétée pour la quantification de chaque nouveau vecteur d'entrée.

L'élimination du chemin ayant la plus grande distorsion parmi les deux chemins arrivant à un même nœud, spécifique à l'algorithme de Viterbi, ne compromet pas l'optimalité de l'algorithme dans le cas des treillis utilisés pour le codage de sources sans mémoire. Ceci n'est plus vrai si la procédure de codage par treillis incorpore de la prédiction, car dans ce cas-là le signal codé futur dépend de l'excitation

[1] Du point de vue de la distorsion due au codage.

du codeur jusqu'à l'instant courant, et par conséquent l'excitation engendrant la moindre distorsion jusqu'à l'instant courant peut être à long terme moins bonne que celle correspondant au chemin qu'on élimine. Pourtant, pour limiter la complexité, la procédure de Viterbi est appliquée aussi dans le cas de la structure prédictive.

Pour limiter le retard de codage, les paramètres de l'excitation doivent être régulièrement transmis au décodeur. Pour ce faire, le meilleur chemin jusqu'au vecteur courant n est identifié et les paramètres correspondants à ce chemin sont transmis, commençant par le vecteur suivant le dernier vecteur transmis auparavant et finissant avec le vecteur $n - L$. Les chemins qui ne sont pas identiques avec le meilleur chemin jusqu'au vecteur $n - L$ sont éliminés (cette opération est appelée le test d'ambiguïté - "ambiguity check"). On peut observer que plus le "retard de décision" L est grand, moins il y a de chemins éliminés par le test d'ambiguïté, et donc avec un retard plus grand on peut espérer une meilleure performance. En pratique, un compromis doit être fait entre la performance et le retard du codeur.

3. ETIQUETAGE DES BRANCHES DU TREILLIS

L'étiquetage des branches du treillis pour la TCQ utilise les règles de partition d'ensemble d'Ungerboeck. Pour la modulation codée par treillis, l'étiquetage des branches a pour but de rendre les deux suites de symboles-canal les plus proches correspondantes à des chemins distincts dans le treillis aussi éloignées que possible (en termes de distance euclidienne). Ainsi on réduit la probabilité des erreurs, qui apparaissent lorsque la séquence-canal reçue au décodeur (la séquence transmise auquelle se rajoute le bruit blanc gaussien introduit par le canal) est plus proche d'une séquence-canal associée à un chemin dans le treillis autre que celui suivi par le codeur.

Cette méthode d'étiquetage des branches peut être utilisée aussi dans le cas de la quantification codée par treillis. On peut regarder la TCQ et la TCVQ comme une quantification vectorielle où le dictionnaire contient des vecteurs très longs, composés de toutes les séquences permises par le treillis de scalaires (pour la TCQ) ou vecteurs (pour la TCVQ) issus du dictionnaire du treillis. Utilisant les règles d'Ungerboeck pour étiqueter les branches du treillis, on maximise la distance minimale entre deux vecteurs de ce dictionnaire. Donc, il n'existe pas des séquences de bits différentes spécifiant des excitations similaires; les bits disponibles sont utilisés efficacement.

La méthode de partition d'ensemble d'Ungerboeck commence par la constellation de modulation (pour la TCVQ cela correspond au dictionnaire du treillis) qui est partitionnée en deux sous-ensembles ayant le même nombre d'éléments, puis chacun de ces deux sous-ensembles est partitionné à son tour en deux, etc. Chaque partition est faite dans le but de maximiser la distance Δ_i entre les deux symboles les plus proches contenus dans chaque sous-ensemble résultant.

Dans le cas du TCVQ CELP nous avons utilisé l'approche suivante pour peupler le treillis:

1. Optimisation du dictionnaire du TCVQ CELP pour le codeur VQ CELP correspondant.
2. Partition du dictionnaire ainsi obtenu.

L'optimisation du dictionnaire du treillis pour le codeur VQ CELP correspondant est faite utilisant la méthode d'optimisation en "boucle fermée" de [14], dans le but de minimiser le bruit de codage pondéré perceptuellement.

Suivant les notations d'Ungerboeck, le dictionnaire optimisé A_0 (de dimension 2^{m+1}) est partitionné en deux sous-ensembles de dimension 2^m, B_0 et B_1; chacun de ces sous-ensembles est ensuite partitionné en deux. On obtient quatre sous-ensembles de dimension 2^{m-1}, C_0 et C_2 et respectivement C_1 et C_3. Chaque partition est faite de manière à maximiser la distance minimale à l'intérieur des sous-ensembles qui en résultent. Par exemple, la partition de A_0 en B_0 et B_1 doit maximiser[2]

$$\Delta_{min} = \min(\min_{x,y \in B_0} d(x,y), \min_{x,y \in B_1} d(x,y)) \quad (4)$$

La partition optimale du dictionnaire suivant les règles d'Ungerboeck est une tâche difficile en termes de complexité car le dictionnaire du treillis n'a pas de structure ou de symétrie bien précise, comme c'était le cas avec les constellations de modulation. La solution directe serait d'évaluer Δ_{min} pour toutes les partitions possibles de A_0 dans deux sous-ensembles de même dimension. Pourtant, quand m augmente, cette méthode devient rapidement très coûteuse en calculs; par exemple, pour $m = 5$, il s'agit de choisir une partition parmi C_{64}^{32}, qui est de l'ordre de 10^{18}. Nous proposons ici deux algorithmes sous-optimaux, que nous illustrons pour la partition de A_0 en B_0 et B_1.

ALGORITHME 1[3]:

1. Trouve c_0^0 et c_1^0 dans A_0 minimisant $d(c_0^0, c_1^0)$; i=1.
2. B_0=$\{c_0^0\}$ et B_1=$\{c_1^0\}$; $A_0 = A_0 \setminus \{c_0, c_1\}$.
3. Trouve c_0^i dans A_0 maximisant $\min|_{y \in B_0} d(c_0^i, y)$; B_0=$B_0 \cup \{c_0^i\}$ et $A_0 = A_0 \setminus \{c_0^i\}$.
4. Trouve c_1^i dans A_0 maximisant $\min|_{y \in B_1} d(c_1^i, y)$; B_1=$B_1 \cup \{c_1^i\}$ et $A_0 = A_0 \setminus \{c_1^i\}$.
5. Si $A_0 \neq \emptyset$, alors $i = i + 1$ et puis retourne à 3. Sinon **fin**.

ALGORITHME 2:

1. Commence par une partition ad-hoc de A_0, $A_0 = B_0 \cup B_1$.
2. S'il existe une paire de vecteurs $c_0 \in B_0$ et $c_1 \in B_1$ tels que:
   a) $\min|_{y \in B_0 \setminus c_0} d(c_0, y) < \min|_{y \in B_1 \setminus c_1} d(c_0, y)$ et
   b) $\min|_{y \in B_1 \setminus c_1} d(c_1, y) < \min|_{y \in B_0 \setminus c_0} d(c_1, y)$
   alors $B_0 = (B_0 \setminus \{c_0\}) \cup \{c_1\}$ et $B_1 = (B_1 \setminus \{c_1\}) \cup \{c_0\}$. Sinon **fin**.
3. Retourne à 2.

[2] Ici $d(x,y)$ est la distance euclidienne entre les vecteurs x et y.

[3] Le symbole "\" signifie ici l'opération de soustraction d'ensemble.

Dans l'algorithme 1 les sous-ensembles B_0 et B_1 sont construits graduellement. A chaque itération chacun de ces deux sous-ensembles se voit attribuer un vecteur de A_0, choisi de façon à maximiser la distance minimum entre ce nouveau vecteur et les vecteurs qui se trouvent déjà alloués à B_0 ou respectivement B_1. Nous avons observé que si le dictionnaire du treillis n'est pas normalisé (si les vecteurs du dictionnaire n'ont pas la même norme), après application de l'algorithme 1, les normes des vecteurs se trouvant dans chacun des sous-ensembles résultants couvrent approximativement le même domaine (figure 2). Nous avons constaté également qu'un dictionnaire non-normalisé donne de meilleurs résultats quand le gain d'excitation est quantifié avec un nombre réduit de bits. L'équivalence des différents sous-ensembles du point de vue de l'energie des vecteurs qu'ils contiennent signifie que le treillis sera équilibré, dans ce sens que pour le codage d'un vecteur "cible" donné, il y aura de bons candidats vecteurs d'excitation dans chaque sous-ensemble.

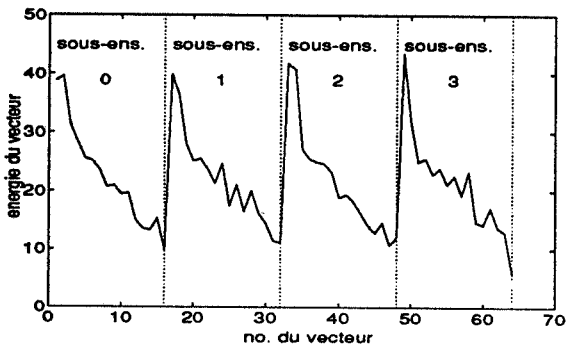


Figure 2. L'énergie des vecteurs se trouvant dans les quatre sous-ensembles du dictionnaire du treillis obtenus après application de l'algorithme 1 (le dictionnaire du treillis contient ici 64 vecteurs).

L'algorithme 2 part d'une partition ad-hoc de A_0 et il cherche des paires de vecteurs $c_0 \in B_0$ et $c_1 \in B_1$ tels que leur permutation engendre une croissance de la distance minimum inter-vecteurs aussi bien à l'intérieur de B_0 qu'à l'intérieur de B_1. Par conséquent, l'algorithme 2 mène à un "minimum local" de Δ_{min} (4), dans la mesure où l'on admet seulement des permutations entre deux éléments de B_0 et respectivement B_1.

Nous avons utilisé ces deux algorithmes pour partitionner le dictionnaire du treillis de la manière suivante. Premièrement, on a appliqué l'algorithme 1 plusieures fois (en partant du dictionnaire optimisé auparavant) et on a gardé la partition qui, utilisée dans le codeur TCVQ CELP, donnait le meilleur résultat du point de vue de la qualité du signal codé[4]. Deuxièmement, on a appliqué l'algorithme 2 à cette partition.

[4]Nous avons apprécié la qualité du signal codé en utilisant le rapport signal/bruit segmental et des tests subjectifs de la qualité perceptuelle

4. COMPLEXITÉ DU TCVQ CELP

Dans le TCVQ CELP, pour chaque chemin survivant, on doit mémoriser:

a) les informations permettant d'effectuer la prédiction à court et long terme:

- le signal d'excitation du passé utilisé pour la prédiction à long terme.
- les mémoires du filtre de synthèse et du filtre de pondération perceptuelle.

b) les paramètres de l'excitation, qui pourront être transmis par la suite.

Si le filtre de synthèse est actualisé à partir du signal codé (pour des codeurs "backward-adaptive"), alors ce signal doit être aussi mémorisé pour chaque chemin.

Le nombre de chemins survivants - les chemins qui restent après élimination par l'algorithme de Viterbi et par le test d'ambiguïté - croît avec le retard de décision: il y a un seul chemin survivant pour un retard de décision 0, 2 pour un retard 1, 4 pour un retard 2, etc. Cette croissance est limitée dans la mesure où le nombre de chemins survivants ne peut pas dépasser le nombre d'états du treillis, N.

Jusqu'au vecteur $n-L$, où n est l'index du vecteur courant et L est le retard de décision en nombre de vecteurs, tous les chemins survivants coïncident, et donc l'on doit mémoriser séparément seulement les informations qui sont plus récentes que le vecteur $n-L$.

Si le filtre de synthèse est actualisé à partir du signal d'entrée (adaptation "forward"), il est de ce fait le même pour tous les chemins. Dans ce cas la recherche du meilleur vecteur d'excitation dans le TCVQ CELP implique seulement une fois le filtrage de chaque vecteur du dictionnaire du treillis et le calcul de l'énergie de ce vecteur. Cela signifie que la recherche du meilleur vecteur d'excitation stochastique est environ deux fois plus complexe pour le TCVQ CELP par rapport au VQ CELP, quel que soit le nombre d'états du treillis N et le retard de décision L[5]. La complexité de la recherche du meilleur retard et gain pour la prédiction à long terme (LTP) peut elle aussi être réduite dans le TCVQ CELP "forward", utilisant le fait que le signal d'excitation est commun à tous les chemins survivants jusqu'au vecteur $n-L$.

Pour le TCVQ CELP "backward" à chaque chemin survivant correspond un filtre de synthèse différent, et donc les dictionnaires filtrés sont différents en fonction du chemin. Une réduction de complexité n'est plus possible dans ce cas. La table 1 présente le rapport entre la complexité des routines du TCVQ CELP consommant le plus de temps-processeur et les routines correspondantes du VQ CELP.

5. PERFORMANCE DU TCVQ CELP

La figure 3 présente le rapport signal/bruit segmental obtenu avec un codeur TCVQ CELP à 8 kbit/s utilisant un treillis à 4 états, en fonction du retard de

[5]On considère que le poids des calculs de produits scalaires est faible dans le volume total des calculs liés à la recherche du meilleur vecteur d'excitation stochastique.

Routine	Forward	Backward
Calcul paramètres LTP	N	N
Recherche excit. stochast.	2	N
Filtrages	N	N

Table 1. Le rapport entre la complexité de trois routines TCVQ CELP et celle des routines VQ CELP correspondantes.

décision. Pour comparaison, on a représenté aussi le rapport signal/bruit segmental obtenu avec le même codeur CELP à excitation codée par VQ. Nous utilisons un codeur CELP "backward" avec des vecteurs d'excitation de 12 échantillons. Les paramètres du prédicteur à long terme sont actualisés tous les 24 échantillons. Le comportement de ce codeur est bien caractérisé par les courbes de la figure 3. Du point de vue du rapport signal/bruit segmental, la performance du codeur CELP excité par TCVQ utilisant 5 bits pour coder chaque vecteur est pratiquement équivalente à celle du même codeur CELP excité par VQ utilisant 6 bits/vecteur (le débit utilisé pour coder les autres paramètres transmis reste le même). Nos tests subjectifs ont montré que cette équivalence est valable aussi du point de vue de la qualité perceptuelle du signal codé.

Cela signifie que dans le cas de ce codeur, l'utilisation de la TCVQ à la place de la VQ permet d'économiser un douzième du débit disponible, tout en gardant la même qualité. Les bits économisés peuvent être utilisés pour améliorer la qualité du codeur TCVQ CELP.

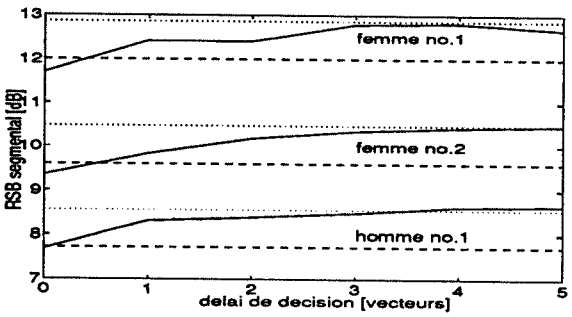


Figure 3. La performance du codeur TCVQ CELP à 5 bits/vecteur pour le codage de trois doubles phrases d'environ 6 secondes chacune. Les lignes interrompue/pointillée représentent la performance du même codeur CELP excité par VQ à 5 bits/6 bits par vecteur.

6. CONCLUSION

Nous utilisons la TCVQ [1] [2] pour le codage de l'excitation d'un codeur CELP. Nos simulations montrent que pour le même débit d'information, le codeur CELP excité par TCVQ est meilleur que le même codeur CELP excité par VQ. La complexité du codeur TCVQ CELP reste raisonnable.

REFERENCES

[1] Malone, K.T., Fischer, T.R., "Trellis-Searched Adaptive Predictive Coding of Speech", IEEE Trans. on Speech and Audio Proc., vol. 1, no. 2, April 1993, pp. 196-206.

[2] Fischer, T.R., Marcellin, M.W., Wang, M., "Trellis Coded Vector Quantization", IEEE Trans. Inform. Theory, vol 37, no. 6, Nov. 1991, pp. 1551-1566.

[3] Marcellin, M.W., Fischer, T.R., "Trellis Coded Quantization of Memoryless and Gauss-Markov Sources", IEEE Trans. Commun.,vol. 38, no.1, Jan. 1990, pp.82-93.

[4] Marcellin, M.W., Fischer, T.R., Gibson, J.D., "Predictive Trellis Quantization of Speech", IEEE Trans. Acoust. Speech Signal Proc., vol. 38, no. 1, Jan. 1990, pp. 46-55.

[5] Schroeder, M.R., Atal, B.S., "Code Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates", Proc. ICASSP 1985, pp. 973-940.

[6] Ungerboeck, G., "Channel coding with multilevel/phase signals", IEEE Trans. Inform. Theory, IT-28, pp. 55-67, Jan. 1982.

[7] Forney, G.D., "The Viterbi Algorithm"(Invited Paper) Proc. IEEE, vol. 61, pp.268-278, Mar. 1973.

[8] Ayanoğlu, E., Gray, R.M., "The Design of Predictive Trellis Waveform Coders Using the Generalized Lloyd Algorithm", IEEE Trans. Comm., COM-34, no.11, Nov. 1986, pp. 1073-1080.

[9] Foster, J., Gray, R.M., Dunham, M.O., "Finite-State Vector Quantization for Waveform Coding", IEEE Trans. Information Theory, IT-31, no. 3, May 1985, pp. 348-359.

[10] Mano, K., Moriya, T., "4.8 kbit/s Delayed Decision CELP Coder Using Tree Coding", Proc. ICASSP 1990, pp. 21-24.

[11] Shoham, Y., "Low Delay Coding of Wideband Speech at 32 kbps Using Tree Structures", in "Speech and Audio Coding for Wireless and Network Applications", Kluwer Academic Publishers, 1993.

[12] Veeneman, D.E., Mazor, B., "An Efficient Structure and Search Strategy for Stochastic Coding at 8 kb/s", Proc. ICASSP 1990, pp. 481-484.

[13] Lin, D., "Speech Coding Using Efficient Pseudo-Stochastic Block Codes", Proc. ICASSP 1987, pp. 1354-1357.

[14] Davidson, G., Yong, M., Gersho, A., "Real-Time Vector Excitation Coding of Speech at 4800 bps", Proc. ICASSP 1987, pp. 2189-2192.

LOCALISATION DE SOURCES VIRTUELLES PAR SYNTHESE BINAURALE. SIMULATION D'UNE ECOUTE STEREOPHONIQUE.

Martine **Marin**, André **Gilloire**, CNET, route de Trégastel, BP 40, 22301 Lannion
Jean-François **Lacoume**, Olivier **Warusfel**, IRCAM, 1 place Stravinsky, 75004 Paris
Jean-Marc **Jot**, Espaces Nouveaux, 56 bd Davout, 75020 Paris.

Résumé : Les techniques de reproduction binaurale permettent de reconstruire un champ acoustique réaliste pour un auditeur muni d'un casque ou d'une paire d'enceintes. Ces techniques font appel aux fonctions de transfert de l'oreille externe (ou HRTF), propre à chaque auditeur, et sont utilisées pour coder l'information de localisation liée à un événement dans l'espace. L'implémentation du filtrage binaural en temps réel sur un système de simulation, est réalisé à l'aide de techniques de modélisation paramétriques. Des tests psycho-acoustiques ont été réalisés pour valider un tel système, en comparant les jugements de localisation obtenus en situation d'écoute réelle d'un dispositif stéréophonique et ceux obtenus par simulation sur casque.

1. INTRODUCTION

Le développement futur des systèmes de téléconférence est conditionné par l'amélioration de la qualité sonore de ces systèmes. En particulier il est souhaitable de se rapprocher des conditions naturelles d'écoute en restituant la localisation auditive des locuteurs de la salle distante.

Pour apporter des solutions techniques appropriées, il est primordial de répondre aux exigences psycho-acoustiques régissant notre perception auditive. Cela implique de connaître et de contrôler ces paramètres subjectifs relatifs à la localisation.

En intégrant ces informations dans un environnement de simulation psycho-expérimental, reproduisant diverses configurations acoustiques de visioconférence, on peut, d'une part, définir les différents attributs perceptifs pertinents pour la téléconférence et, d'autre part, étudier l'influence des paramètres objectifs liés aux systèmes de prise ou de restitution du son sur la localisation subjective.

Dans le cas d'une reproduction sur casque (ou sur une paire d'enceintes), la localisation est restituée artificiellement en temps réel en utilisant les techniques de synthèse binaurale nécessitant l'implémentation des HRTF (Head Related Transfert Function) par filtrage numérique. Les HRTF, qui représentent le filtrage opéré entre une source dans une direction donnée et l'intérieur du conduit auditif, sont propres à chaque individu; de telles techniques nécessitent donc la mesure des HRTF pour chaque sujet.

La validation de cet environnement de test, a été réalisée en comparant les jugements de localisation obtenus en situation d'écoute réelle d'un dispositif stéréophonique conventionnel et ceux obtenus par simulation sur casque de ce même dispositif.

2. PERCEPTION DE L'ESPACE SONORE

La localisation d'un événement sonore par le système auditif est déterminée à partir du champ acoustique créé au niveau du tympan [1]. Les principaux paramètres contribuant à la localisation sont :

- la source sonore (sons purs ou complexes, impulsions);
- l'environnement acoustique (salle, champ libre);
- l'influence du haut du corps de l'auditeur sur l'onde incidente (plan vertical);
- la tête (mouvements, ombrages);
- l'oreille externe et plus spécifiquement le pavillon;
- les stimuli visuels et l'expérience auditive.

La position d'une source est repérée dans l'espace tridimensionnel par son azimut φ, son élévation ε, et sa distance r.

L'azimut φ est déterminé essentiellement grâce à deux indices obtenus en comparant le champ sonore parvenant à chaque oreille. Les différences interaurales d'intensité (ILD) sont négligeables en basse fréquence et augmentent avec la fréquence par diffraction autour du buste et de la tête. Les différences interaura-

les de temps (ITD) regroupent les différences de phase, utilisables au dessous de 1500 Hz et les différences de temps d'arrivée de la structure globale, qui prennent le relais des différences de phase. Dans le cas d'une source frontale ($\varphi = 0$) ou arrière ($\varphi = 180$), ces différences interaurales sont nulles et de fréquentes inversions avant-arrière apparaissent. De telles ambiguités sont levées par des petits mouvements de tête.

Dans le plan vertical, le rôle du pavillon devient primordial et le codage est essentiellement spectral, d'ou l'importance de la nature des signaux. La présence des bandes directives impose une direction apparente de la source, déterminée par les zones spectrales où le signal possède un maximum d'énergie.

Quelle que soit la position spatiale de la source, le signal subit une altération spectrale par le pavillon. Celui-ci agit comme un filtre linéaire et code les informations spatiales du champ sonore en informations spectrales et temporelles.

Pour déterminer la distance r de la source à l'auditeur, le système auditif se base sur des variations d'intensité (doublement de la distance subjective pour une réduction de niveau de 20 dB) et spectrales (absorption plus importante des hautes fréquences par l'air), apparaissant lors du déplacement de cette source. Dans le cas d'une salle, le rapport champ direct / champ réverbéré est un indice supplémentaire permettant de juger du positionnement absolu de la source.

3. FONCTION DE TRANSFERT DE L'OREILLE EXTERNE

3.1. Définition

Au cours de sa propagation entre la source et le tympan, l'onde sonore subit un filtrage caractéristique d'une direction donnée, se traduisant par des différences temporelles et d'intensité et des modifications spectrales. La fonction de transfert de l'oreille externe ou HRTF, qui représente la fonction de transfert du canal acoustique entre la source et le tympan, code simultanément ces trois indices.

Dans l'optique d'une simulation binaurale de localisation de sources sonores, il faut mesurer les HRTF pour différents azimuts et élévations de la source dans des conditions de champ libre. On peut montrer qu'il suffit de mesurer le signal à l'entrée du canal auditif pour avoir accès à toutes les caractéristiques spatiales du signal car la dépendance des HRTF sur la direction n'est affectée ni par la forme du canal auditif ni par l'impédance du tympan [1].

3.2. Mesures des HRTF

En raison de l'importance des différences inter-individuelles des HRTF, nous avons constitué notre propre base de données de HRTF en les mesurant sur une vingtaine de sujets au CNET et à l'IRCAM (fig. 1).

Pour chaque sujet, 48 mesures ont été réalisées: l'ensemble du plan horizontal a été exploré par pas de 15° pour les élévations ε=0° et ε=30°. Le sujet était placé en chambre sourde sur un siège pivotant, à 2 m de la source. Les récepteurs étaient des microphones miniatures à électret, directement insérés dans le conduit auditif du sujet. Les réponses impulsionnelles ont été mesurées à 48 kHz à partir d'un haut-parleur à deux voies concentriques, excité par des séquences pseudo-aléatoires [2].

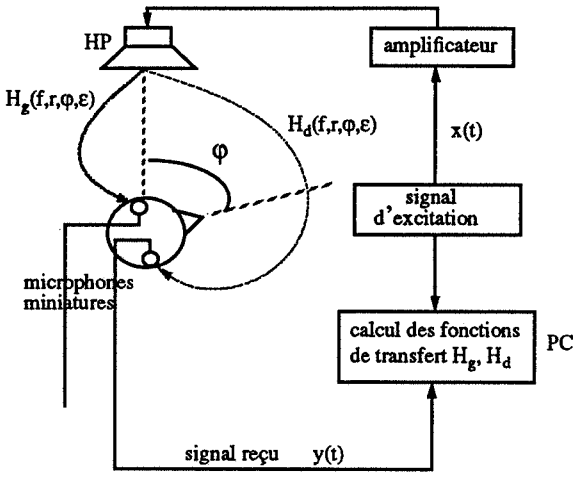


Fig. 1 : mesure des fonctions de transfert H_g et H_d.

Afin d'obtenir une mesure correcte, il faut s'affranchir des systèmes électro-acoustiques qui entachent la mesure. Notamment pour une reproduction ultérieure au casque, il est nécessaire de mesurer, en plus pour chaque oreille, la fonction de transfert entre l'écouteur et le microphone placé à la même position dans le conduit que lors des mesures précédentes afin d'égaliser le casque par inversion de cette fonction de transfert.

4. SYNTHESE DE LA LOCALISATION

4.1. Principe

La synthèse binaurale repose sur l'hypothèse qu'une reproduction fidèle sera obtenue si on reconstruit le signal de pression capté par les tympans. La simulation d'un environnement acoustique donné peut être réalisée par convolution avec les réponses impulsionnelles mesurées au niveau du tympan dans ce même environnement. En ce qui concerne la localisation, elle est restituée en filtrant le signal par les HRTF droite et gauche, mesurées en champ libre dans le cas d'un événement sonore placé dans la position désirée par rapport à l'auditeur [1].

$$y_g(t) = [h_g(t) * x(t)] * h_{cg}^{-1}(t)$$

$$y_d(t) = [h_d(t) * x(t)] * h_{cd}^{-1}(t)$$

où h(t) et h_c(t) sont les réponses impulsionnelles respectivement des HRTF correspondantes à la posi-

tion désirée de la source et de la fonction de transfert du casque. Les signaux issus de cette convolution sont alors directement utilisables pour une écoute au casque au même titre que les enregistrements par tête artificielle. Une reproduction sur haut-parleur peut aussi être envisagée mais il faut alors introduire un filtrage supplémentaire qui annule les chemins croisés entre un haut-parleur et l'oreille opposée (crosstalk cancellation) [3]. De plus, l'utilisation du casque comme moyen de reproduction pour les expériences psychoacoustiques permet un meilleur contrôle des stimuli et ne requiert pas d'environnement acoustique de simulation spécifique.

4.2. Modélisation et implémentation

L'implémentation en temps réel des filtres binauraux sous la forme de filtres FIR par convolution directe avec les coefficients de la réponse impulsionnelle nécessite une puissance de calcul élevée. Afin de diminuer le coût de calcul, il est possible de réduire l'information contenue dans les HRTF sans perdre pour autant les caractéristiques indispensables à la localisation. Pour cela une modélisation paramétrique a été effectuée sur chaque réponse impulsionnelle, fournissant un filtre récursif équivalent. La modélisation s'appuie sur deux propriétés [4]:

- une propriété générale des filtres rationnels stables, qui permet de décomposer tout système en un filtre à phase minimale associé, en série, avec un filtre passe-tout modélisant l'excès de phase;
- une propriété particulière des HRTF, pour lesquelles l'excès de phase est linéaire en fonction de la fréquence, rendant le filtre passe-tout équivalent à un retard pur (fig. 2).

Avant de synthétiser le filtre IIR à phase minimale, on effectue préalablement un lissage perceptif du module de la réponse fréquentielle de la HRTF selon les bandes critiques de l'oreille, afin d'éliminer la structure fine des réponses tout en respectant la résolution fréquentielle de l'oreille [5]. Un filtre IIR d'ordre 20 pour chaque filtre binaural suffit à reproduire correctement les caractéristiques de localisation (fig. 3).

L'égalisation du casque est réalisée par la mise en série des filtres binauraux et du filtre inverse du casque, lui aussi réalisée sous la forme de filtres IIR d'ordre 20 (fig. 4).

La simulation en temps réel est réalisée sur une station de travail Next Cube, munie d'une carte de traitement de signal ISPW (deux micro-processeurs Intel i860) travaillant à la fréquence d'échantillonnage de

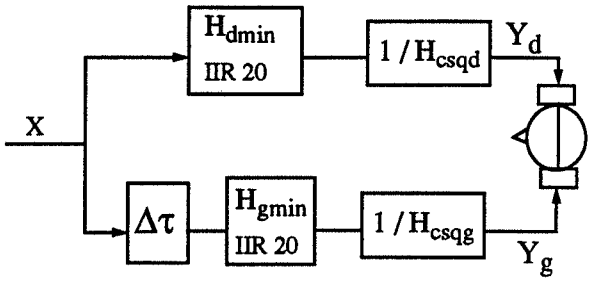


Fig. 4 : synthèse binaurale avec implémentation des filtres binauraux à l'aide de filtres IIR d'ordre 20 et égalisation du casque. $\Delta\tau$ représente le retard interaural (cas d' une source à droite).

44.1 kHz. Les modules de traitement de signal et l'interface de commande sont développés grâce au logiciel Max (environnement graphique orienté objet).

5. VALIDATION DE L'ENVIRONNEMENT DE SIMULATION

Pour obtenir des indications sur les biais éventuels apportés par l'environnement de simulation, nous avons délaissé l'approche globale qui consisterait à comparer les situations d'écoute réelles et ces mêmes situations simulées et à vérifier qu'elles sont indiscernables. Les différentes étapes de validation sont menées selon une approche structurée consistant à isoler le paramètre de localisation, et d'étudier si cette dimension est traduite de manière fidèle par la simulation. Lors de la validation décrite ci-dessous il s'agit de comparer les jugements de localisation obtenus en situation d'écoute stéréophonique réelle sur haut-parleurs à ceux obtenus par simulation sur casque [6].

Comme l'indique la fig. 5 on simule un dispositif

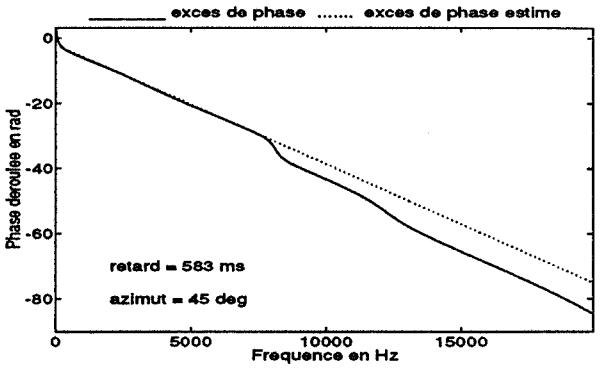


Fig. 2 : modélisation de la phase.

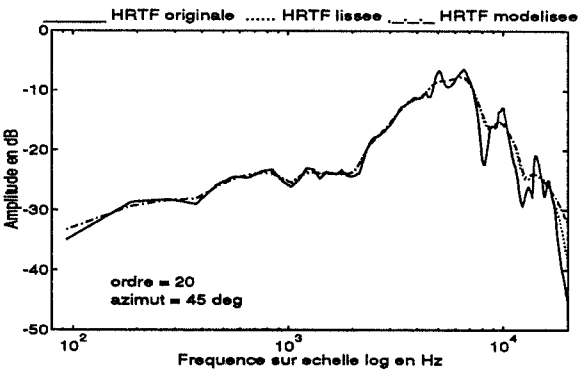


Fig. 3 : modélisation de l'amplitude.

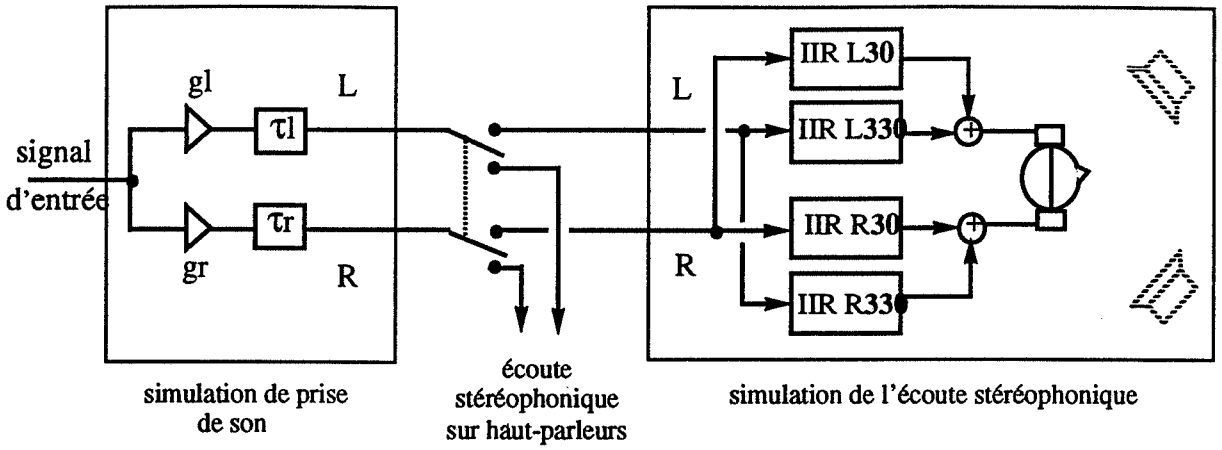


Fig. 5 : simulation d'un dispositif de reproduction stéréophonique en plaçant deux sources virtuelles respectivement à +30° et -30° par rapport au sujet, représentant les haut-parleurs virtuels.

de reproduction stéréophonique standard dans un milieu anéchoïque, en reproduisant deux sources virtuelles placées respectivement à +30° et -30° par rapport au sujet. Les signaux présentés sur chaque oreille sont donc la somme des contributions des deux hauts-parleurs, chacune de ces contributions étant reproduite en filtrant le signal gauche (respectivement droit) par les HRTF gauche et droite correspondantes à la position +30° (respectivement -30°).

La prise de son est simulée lors de l'expérience en introduisant des différences de niveau entre les deux canaux (couple XY), auxquelles s'ajoute un décalage temporel dans le cas d'un couple non coïncident (couple AB ORTF).

Les signaux de tests sont des EMDIFS [7], répartis de 125 à 8000 Hz, et deux signaux de parole anéchoïques (une voix masculine et une voix féminine issues d'un corpus de phrases phonétiquement équilibrées). Les EMDIFS sont des sinusoïdes avec une enveloppe temporelle reproduisant les courbes de masquage fréquentiel de l'oreille. Elles sont donc perçues comme des fréquences pures.

La variation d'angle minimum audible étant de 3°, la source sonore est enregistrée par pas de 5° entre 0° et le 1/2 angle de prise de son, angle à partir duquel la source virtuelle semble provenir de l'une des deux enceintes. Cet angle varie selon le couple stéreophonique utilisé, chaque couple possèdant sa propre courbe de latéralisation (angle apparent de restitution en fonction de l'angle de la source enregistrée), en général non linéaire. Ainsi nous avons utilisé un 1/2 angle de prise de son de 90° pour le couple XY et de 45° pour le couple AB.

L'ensemble des stimuli est présenté dans un ordre aléatoire pour réduire les effets psycho-acoustiques de persistance et éviter d'orienter les réponses comme cela se produit lors d'une progression ordonnée en azimut. Il est demandé à l'auditeur d'évaluer directement l'azimut de la source virtuelle (pas de jugement sur la distance appparente) à partir de repères visuels, disposés tous les 5° entre les deux enceintes, qui restent visibles dans les deux situations d'écoute, réelle et simulée. L'auditeur enregistre ses résultats à partir d'un interface de controle placé devant lui.

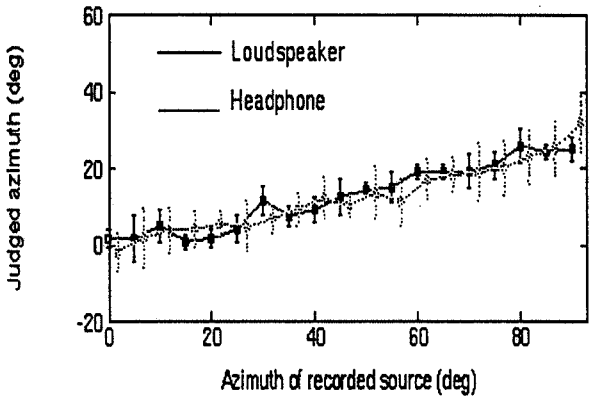


Fig. 6 : comparaison des jugements de localisation (voix féminine, couple XY).

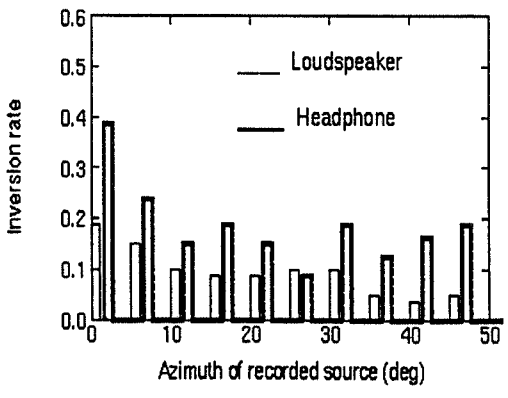


Fig. 7 : comparaison du taux d'inversion avant-arrière (couple AB ORTF).

Dans un premier temps, l'environnement acoustique choisi est anéchoïque. Dans une étape ultérieure, l'expérience sera reproduite dans un environnement de type salle de visioconférence, en synthétisant en plus la distribution spatiale et temporelle d'un effet de salle, grâce au prototype "Spatialisateur" développé actuellement par l'IRCAM et Espaces Nouveaux [8].

6. RESULTATS ET CONCLUSION

La restitution de la latéralisation a été analysée du point de vue des biais éventuels et de la dispersion des résultats entre les sujets, mais aussi du point de vue des inversions avant-arrière, phénomène qui consiste à entendre la source à un azimut symétrique par rapport à l'axe des oreilles de l'azimut théorique (cf §2). Les moyennes et les écarts types de localisation ont été calculés après correction des réponses localisées vers l'arrière.

Pour la latéralisation du couple XY, il apparait une bonne adéquation entre la simulation et la réalité. Les biais sont faibles et les dispersions comparables dans les deux situations (fig. 6). On retrouve les propriétés de latéralisation du couple XY en fonction de la fréquence telles qu'elles sont décrites, par exemple, dans [7]. Les taux d'inversions reflètent bien d'une part, la courbe de localisation des différentes fréquences de Blauert (les fréquences autour de 1000 Hz sont plutôt localisées vers l'arrière) et d'autre part, l'ambiguité engendrée par la position frontale (fig. 7). Nous avons remarqué qualitativement que la source apparente dans le cas des signaux de parole, était plus large et plus instable que dans le cas des EMDIFS, phénomène qui peut s'expliquer par la sensibilité fréquentielle de la localisation vue précédemment.

En ce qui concerne le couple AB ORTF, on constate dans l'expérience sur haut-parleurs une latéralisation plus fidèle à l'angle de prise de son que pour le couple XY. Cette propriété est liée à l'intervention d'indices temporels fournis par les couples non coïncidents. On note cependant une dispersion plus large des réponses des sujets qui peut s'interpréter par le fait que la relation entre indices temporels et indices énergétiques réalisée par le couple AB n'est pas celle de l'écoute naturelle. La concommitance de deux types d'indices rend les performances de localisation plus dépendantes de l'individu et de sa position par rapport aux haut-parleurs. On constate une dispersion comparable lors de l'expérience sur écouteurs (notons que dans ce cas, les différences de positionnement des sujets par rapport aux haut-parleurs se retrouvent dans l'étape de mesure des HRTF).

Pour le couple AB simulé sur écouteurs, l'analyse montre cependant une surestimation de la latéralisation, ainsi qu'une forte dispersion, pour la voix féminine et les signaux centrés sur 1 kHz. Ce phénomène n'apparait que pour de fortes latéralisations, et peut s'interpréter par l'impossibilité, pour le sujet, de préciser son jugement de localisation par des mouvements de la tête. Dans ce cas, certains sujets ont eu tendance à indiquer une latéralisation maximale (90°). Par ailleurs, comme l'illustre la fig. 7, la fréquence des inversions avant-arrière, non négligeable en écoute anéchoïque réelle, est plus élevée encore pour la simulation sur écouteurs. L'analyse montre que cette différence entre simulation binaurale et réalité est plus grande pour la direction frontale et les latéralisations fortes, stimuli susceptibles de provoquer un mouvement réflexe du sujet pour conforter son jugement auditif.

Les conclusions de cette expérience suggèrent donc que les performances du système de simulation, dans la configuration étudiée, seraient améliorées par l'utilisation d'un capteur de position de la tête, permettant un asservissement du traitement binaural relativement à l'orientation du sujet.

BIBLIOGRAPHIE :

[1] BLAUERT J., "Spatial hearing : the psychophysics of human sound localization", Cambridge MIT Press 1983.

[2] MARIN M., "Etude de la localisation en restitution des sons. Application à l'amélioration des systèmes de téléconférence", rapport CNET/LAA/TSS/CMC, 1993.

[3] COOPER D.H., BAUCK J.L., "Prospects for transaural recording", J. Audio Eng. Soc., vol.37(2), 3-19, 1989.

[4] PONCET F., "Simulation de localisation de sources sonores dans l'espace", rapport Télécom Paris, Dpt Signal, 1992.

[5] SMITH J.,"Techniques for digital filter design and system identification with application to the violin", PhD thesis, CCRMA, Dept of Music, Stanford Univ., June 1983.

[6] LACOUME J.F., "Test d'un environnement de simulation de situations d'écoute binaurale sur des signaux stéréophoniques", rapport CNET/IRCAM, 1993.

[7] HUGONNET C.,"Restitution spatiale des différents systèmes stéréophoniques. Caractérisation objective et subjective", mémoire d'ingénieur CNAM, Paris, 1987.

[8] JOT J.M., WARUSFEL O., KAHLE E., MEIN M., "Binaural concert hall simulation in real time", IEEE digital audio workshop, Mohonk, Oct. 1993.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION AFFICHEE : A3

BASES DE DONNEES ET DE CONNAISSANCES

INVESTIGATION DE PROTOCOLES POUR LA RECONNAISSANCE DES LOCUTEURS FONDEE SUR LE SYSTEME ORPHEE

Jean-Luc LE FLOCH, Claude MONTACIÉ & Marie-José CARATY

LAFORIA-IBP U.R.A. 1095 C.N.R.S.
UNIVERSITÉ P. & M. CURIE - 4, PLACE JUSSIEU
75252 PARIS CEDEX 05 - FRANCE

RÉSUMÉ

Les performances de reconnaissance du locuteur donnent une information sur la capacité des paramètres de protocoles de reconnaissance à discriminer des locuteurs. Les travaux présentés dans cet article sont fondés sur des protocoles qui se sont avérés être robustes au bruit : les protocoles de la serrure vocale Orphée du Laforia (i.e., modèle AR-Vectoriel et distance d'Itakura vectorielle). L'influence de la durée de parole, l'effet des paramètres du modèle AR-Vectoriel et l'importance des segments phonétiques sont étudiés par des expériences en reconnaissance du locuteur indépendamment du texte. A cet effet, une méthode originale a été développée pour déterminer quels sont les phonèmes les plus caractéristiques du locuteur dans ces expériences.

Les résultats montrent que les voyelles, les diphtongues et les nasales sont les classes phonétiques les plus porteuses de l'information caractéristique du locuteur. Une seconde de test composée uniquement de voyelles et de diphtongues permet de reconnaître à 95,6 % un locuteur parmi 630.

Mots-Clés: *reconnaissance du locuteur, modélisation prédictive, segments phonétiques*

INTRODUCTION

La détermination des traits caractéristiques du locuteur est fondamentale pour résoudre deux problèmes: la normalisation du locuteur pour la reconnaissance de la parole et la reconnaissance du locuteur (i.e., identification/vérification).

Les performances de reconnaissance du locuteur donnent une information sur la capacité des paramètres de protocoles de reconnaissance à discriminer des locuteurs. Dans ce but, nous utilisons le système Orphée : la serrure vocale de la porte du Laforia. Orphée est située dans un endroit extrèmement bruyant et réverbérant (i.e., une cage d'escalier en béton). La serrure est utilisée depuis 9 mois par 90 membres du laboratoire sans aucune entrée d'imposteur. Les statistiques d'utilisation sont les suivantes : en moyenne, sur 20 tentatives par jour, on relève une tentative déjouée d'imposteur et 12 entrées réussies. L'analyse des 7 autres tentatives montrent vraisemblablement que ces échecs sont dus à une trop courte durée de parole (i.e., < 2 secondes) ou à de mauvaises conditions d'enregistrement (e.g., bruit intempestif, intensité du signal de parole).

Parmi l'ensemble des protocoles testés en reconnaissance du locuteur, nous présentons les résultats sur deux des principales caractéristiques du signal vocal analysé : sa durée et son contenu phonétique.

1. MODÈLES AR-VECTORIEL ET MESURE INTER-LOCUTEURS

La modélisation AR-Vectoriel (ARV) est un outil classique en traitement du signal multi-dimensionnel. Elle est utilisée ici pour décrire les trajectoires des vecteurs cepstraux d'un signal de parole.

Soit $\{y_n\}$, (n= 1,..., N) une succession de N vecteurs cepstraux d'ordre p. Leur évolution est décrite par un modèle auto-régressif vectoriel d'ordre q.

$$y_n = \sum_{i=1}^{q} A_i \, y_{n-i} + e_n$$

où $\{A_i\}$, (i= 1,..., q) sont les matrices (p, p) et $\{e_n\}$, (n = 1,..., N) est un bruit vectoriel blanc qui a pour matrice de covariance D.

Les coefficients des matrices A_i sont estimés par l'algorithme de Levinson-Whittle-Robinson (cf. Appendice). Le critère à minimiser est la trace de la matrice D. Appliqués à la reconnaissance du locuteur [2,3,4], ces modèles peuvent être interprétés comme une représentation des capacités articulatoires du locuteur. La mesure inter-locuteurs choisie IS_I [3] est basée sur la mesure symétrique semblable à celle d'Itakura.

2. BASE DE DONNÉES

La totalité de la base de données TIMIT [5] a été utilisée pour les expériences. Cette base de données consiste en l'enregistrement de 630 locuteurs coopératifs (i.e., 192 femmes et 438 hommes). Chaque locuteur a prononcé 10 phrases. 2 de ces phrases sont des phrases "dialectales" dites par chaque locuteur, les 8 autres phrases sont différentes pour chaque locuteur. 5 d'entre elles sont des phrases "MIT", les 3 restantes sont des phrases "TI".

Deux ensembles d'apprentissage (App) et deux ensembles de test (Test) sont définis. App1 est composé des cinq phrases "MIT", Test1 des 2 phrases "dialectales" et des 3 phrases "TI". App1 est subdivisé en deux sous-ensembles App2 et Test2 qui vont permettre de choisir de nouveaux protocoles tout en conservant Test1 pour l'évaluation des protocoles. App2 est composé des quatre premières phrases "MIT", et Test2 de la dernière phrase "MIT". Seuls les événements correspondant à un signal de parole sont considérés. La segmentation utilisée dans les expériences est celle fournie avec TIMIT.

3. INFLUENCE DE LA DURÉE DU SIGNAL DE TEST

Les expériences sur l'influence de la durée du signal de test sont menées sur App1 et une partie de Test1. Les modèles ARV de chaque locuteur sont appris sur App1. Le signal de test est obtenu en concaténant les phrases de Test1 jusqu'à concurrence de la durée souhaitée. Pour garantir l'indépendance du contenu phonétique, les phrases sont concaténées dans le sens inverse de leur apparition dans la base de données TIMIT. La dimension p des vecteurs cepstraux varie de 4 à 20 pour différentes valeurs de l'ordre q des modèles ARV. La durée de test varie de 0,5 à 12 secondes.

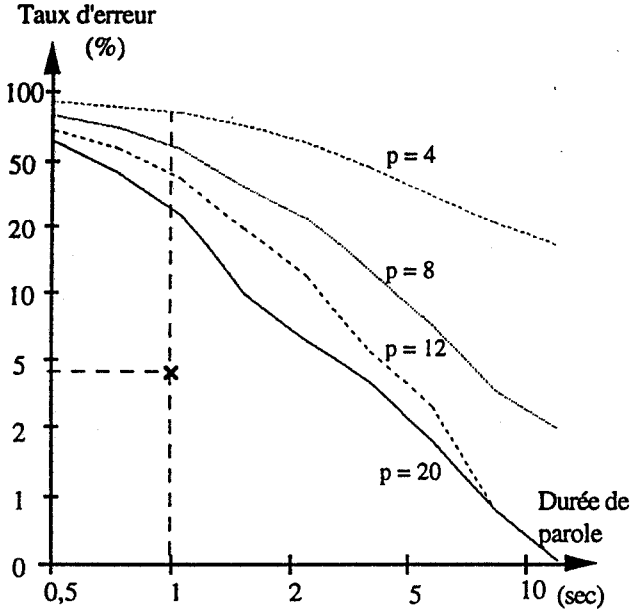


Figure 1. Courbes des erreurs en fonction de la durée du signal de test pour diverses valeurs de p (modèles ARV d'ordre 2).

Nous remarquons que le taux d'erreur de reconnaissance du locuteur est inversement proportionnel à la durée du signal de parole et diminue avec la dimension des vecteurs cepstraux. Dans des études antérieures, nous avions observé que ce taux était proportionnel à la racine carrée du nombre de locuteurs [2, 6]. Cependant, ces résultats sont obtenus à partir d'un maximum de 12 secondes de parole et ne peuvent a priori pas être extrapolés pour une durée d'ordre bien supérieure (e.g., 1 heure).

Ordre du modèle AR	q = 2	q = 3	q = 4	q = 5
p = 4	22,8%	29,2%	30,5%	30,5%
p = 8	53,0%	57,5%	60,5%	60,3%
p = 12	68,7%	69,5%	72,0%	71,7%
p = 16	78,9%	77,5%	—	—
p = 20	81,6%	—	—	—

Tableau 1. Taux de reconnaissance du locuteur, pour une seconde de test et différentes valeurs de p et q.

Dans ce tableau, les résultats de reconnaissance du locuteur sont présentés pour une seconde de signal test. On note que ce taux augmente puis se stabilise avec l'ordre q des modèles ARV pour p inférieur à 16. La diminution des taux en fonction de l'ordre q pour p supérieure à 16 est vraisemblablement due à une quantité insuffisante de vecteurs d'appentissage pour estimer les matrices A_i.

4. RECONNAISSANCE ANALYTIQUE ET SEGMENTALE DU LOCUTEUR

Deux séries d'expériences de reconnaissance (i.e., analytiques et segmentales) sont menées sur Test2. Les modèles ARV de chaque locuteur sont appris sur App2. Dans une reconnaissance "analytique", tout ensemble de q+1 vecteurs cepstraux consécutifs (nécessaires à la prédiction vectorielle) permet une reconnaissance du locuteur. La décision est prise en comparant le $(q+1)^{ème}$ vecteur avec la prédiction réalisée par les modèles ARV d'apprentissage sur les q premiers vecteurs. Deux taux de reconnaissance analytique du locuteur sont calculés par phonème. Le premier est la moyenne des taux de reconnaissance pour chaque ensemble de q+1 vecteurs consécutifs du segment phonétique considéré. Pour le second, seul le tiers central des segments phonétiques est utilisé. Dans une reconnaissance "segmentale", un segment phonétique permet une reconnaissance du locuteur à partir de la distance IS_I entre le modèle ARV du segment et les modèles ARV appris.

Voyelles					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
iy	12,3 %	12,1 %	762	9	7,4 %
ih	15,9 %	16,9 %	718	8	9,6 %
ix	14,5 %	15,7 %	1210	5	3,0 %
eh	17,5 %	18,9 %	539	9	16,5 %
ae	16,1 %	17,5 %	368	13	14,7 %
ah	16,3 %	17,6 %	345	9	17,1 %
ax	14,8 %	16,4 %	476	5	2,9 %
ax-h	6,7 %	8,8 %	69	3	0,0 %
aa	17,3 %	18,4 %	376	12	19,7 %
ao	13,8 %	13,5 %	272	12	14,3 %
uh	13,4 %	14,0 %	50	8	4,0 %
uw	15,3 %	17,2 %	45	10	15,6 %
ux	10,3 %	10,0 %	180	10	5,0 %
er	14,7 %	15,5 %	240	12	14,2 %
axr	11,2 %	12,3 %	422	8	4,0 %
Diphtongues					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
ey	16,1 %	17,4 %	354	13	15,8 %
ay	17,1 %	17,7 %	265	15	21,9 %
oy	14,3 %	15,8 %	27	17	29,6 %
aw	15,3 %	16,8 %	72	16	34,7 %
ow	17,8 %	18,0 %	182	13	14,3 %

Nasales					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
m	13,1 %	13,3 %	563	6	4,4 %
em	10,7 %	10,8 %	16	8	0,0 %
n	13,8 %	13,8 %	981	5	3,5 %
en	12,3 %	13,0 %	111	8	6,3 %
nx	17,4 %	15,1 %	114	3	0,9 %
ng	12,4 %	12,2 %	205	6	2,4 %
eng	18,9 %	18,4 %	4	8	0,0 %
Liquides, Glissées, Aspirées et Silences					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
l	13,0 %	12,9 %	721	6	3,3 %
el	14,7 %	15,9 %	158	9	8,9 %
r	13,2 %	13,6 %	830	5	2,7 %
w	12,8 %	13,2 %	302	6	2,7 %
y	9,5 %	10,6 %	155	5	1,9 %
hh	9,6 %	9,3 %	99	6	0,0 %
hv	8,0 %	9,0 %	64	7	0,0 %
pau	7,9 %	7,0 %	82	17	2,4 %
epi	5,6 %	5,4 %	131	4	0,8 %
Occlusives					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
p	7,4 %	8,0 %	514	4	0,6 %
t	6,3 %	6,1 %	648	5	0,3 %
k	6,4 %	5,9 %	767	5	0,3 %
pcl	7,0 %	6,2 %	504	7	0,0 %
tcl	6,2 %	5,9 %	1014	5	0,1 %
kcl	6,4 %	6,1 %	799	6	0,1 %
q	9,4 %	8,9 %	385	6	1,0 %
b	8,6 %	7,1 %	334	2	0,6 %
d	8,6 %	9,0 %	421	2	0,0 %
g	5,7 %	5,3 %	249	3	1,6 %
bcl	9,3 %	8,9 %	294	6	0,7 %
dcl	8,6 %	7,9 %	677	5	0,3 %
gcl	7,9 %	7,9 %	235	6	0,9 %
dx	9,6 %	9,3 %	233	3	0,9 %
Fricatives et Affriquées					
Ph.	Reconnais. analytique Ph.	Reconnais. analytique Cent. Ph.	Nb.	Durée moy. (csec)	Reconnais. segmentale Ph.
s	6,8 %	6,9 %	1078	11	1,4 %
sh	5,2 %	4,5 %	255	12	0,8 %
th	5,8 %	5,4 %	75	9	0,0 %
f	6,4 %	5,9 %	309	10	0,0 %
z	7,1 %	7,0 %	506	8	0,8 %
zh	3,6 %	3,5 %	30	8	0,0 %
dh	8,1 %	8,5 %	273	3	0,4 %
v	10,6 %	10,2 %	192	6	1,6 %
ch	4,0 %	3,3 %	184	9	1,1 %
jh/j	5,2 %	5,2 %	240	6	1,3 %

Tableau 2. Taux de reconnaissance analytique et segmentale du locuteur par phonème (p = 8, q = 2).

A priori, les phonèmes contenant le moins d'information caractéristique du locuteur sont les phonèmes présentant un faible taux de reconnaissance analytique et segmentale. Les résultats de reconnaissance analytique du locuteur montrent que les phonèmes les plus discriminants, par ordre décroissant, sont : les diphtongues, les voyelles, les nasales/liquides, les glissées, les plosives/ fricatives/affriquées. On remarque que le taux de reconnaissance analytique du locuteur, pour les sons voisés, est généralement plus élevé lorsqu'il est calculé à partir du tiers central des phonèmes. Ceci est vraisemblablement dû à la stabilité du modèle ARV au centre de ces phonèmes.

Les résultats de reconnaissance segmentale sont faibles pour les phonèmes d'une durée de réalisation inférieure à une dizaine de centisecondes. Par contre, les diphtongues correspondent à de bons résultats explicables par leur importante durée de réalisation. Parmi les phonèmes ayant une durée "suffisante"et dont les résultats sont significatifs, la liste de ceux qui correspondent à un taux de reconnaissance faible est la suivante : [uh], [ux], [axr], [s], [sh], [th], [f], [z], [ch].

5. RECONNAISSANCE DU LOCUTEUR FONDÉE SUR DES CLASSES PHONÉTIQUES

Le but des expériences suivantes est de savoir si les modèles ARV appris sur des segments phonétiques sélectionnés permettent une amélioration des performances de reconnaissance du locuteur. Dix classes phonétiques de base sont définies (cf. Tableau 3).

1	Voyelles	iy, ih, ix, eh, ae, ah, ax, ax-h, aa, ao, uh, uw, ux, er, axr
2	Diphtongues	ey, ay, oy, aw, ow
3	Nasales	m, em, n, en, nx, ng, eng
4	Liquides, Glissées	l, el, r, w, y
5	Occlusives	p, t, k, q, b, d, g, bcl, dcl, gcl, dx
6	Fricatives, Affriquées	s, sh, th, f, z, zh, dh, v, ch, jh/j
7	Silence	pau, epi, pcl, tcl, kcl
8	Voyelles extrêmes	iy, ih, ae, aa, ao, uh, uw, ux
9	Voisées	{1}, {2}, {3}, {4}, b, d, g, dx, z, zh, dh, v, jh/j
10	Non Voisées	p, t, k, q, s, sh, th, f, ch, hh, hv

Tableau 3. Description des classes phonétiques à partir des symboles phonétiques utilisés dans la transcription de la base de données TIMIT.

Les expériences sont menées sur des sous-ensembles de Appl et de Test1 correspondant à la concaténation des segments appartenant à la classe phonétique testée. La durée du test choisie est d'une seconde au maximum. Pour chaque classe phonétique sont donnés (cf. Tableau 4) : le taux de reconnaissance du locuteur, la moyenne de la durée totale des segments phonétiques utilisés pour

l'apprentissage, et le pourcentage de locuteurs dont la durée de test est inférieure à une seconde.

Classes phonétiques	1	2	3	4	5
Taux de reconnaissance	94,3	55,2	73,7	47,6	12,5
Durée d'apprent. (csec)	492	125	90	108	163
Durée de test < 1 sec %	0	33,5	78,6	3,5	2,5
Classes phonétiques	6	7	8	9	10
Taux de reconnaissance	36,8	3,5	82,1	92,1	31,0
Durée d'apprent. (csec)	233	114	238	848	255
Durée de test < 1 sec %	0	30	0	0	0
Classes phonétiques	1+2	3+4	1+3	1+2 +3	7+9 +10
Taux de reconnaissance	95,6	87,0	94,6	94,1	80,2
Durée d'apprent. (csec)	617	198	582	672	1206
Durée de test < 1 sec %	0	0,2	0	0	0

Tableau 4. Reconnaissance du locuteur à partir des classes phonétiques sur un test d'une seconde au maximum (p= 20, q= 2).

Les meilleurs résultats sont obtenus pour le regroupement des classes phonétiques voyelles et diphtongues. Une amélioration significative de 14% est obtenue comparativement aux résultats originaux (cf. Fig.1 & Tab.1). Le mauvais résultat des diphtongues est surprenant et ne s'explique pas par une mauvaise estimation des modèles ARV. Par contre, on peut remarquer le bon résultat des nasales compte tenu de la courte durée de l'apprentissage et du test. Les autres résultats sont conformes à ceux obtenus dans les expériences de reconnaissance (analytique et segmentale).

CONCLUSION

Nous avons développé un ensemble de méthodes pour déterminer quels sont les phonèmes les plus caractéristiques du locuteur. Dans ce but, nous avons utilisé les protocoles du système Orphée. Le principal résultat est le suivant : pour la reconnaissance du locuteur, les segments phonétiques à apprendre sont les voyelles et les diphtongues ou les nasales. Ces résultats confirment les précédentes études sur la caractérisation du locuteur [7, 8, 9]. A partir de ces expériences, nous avons montré qu'il est possible de reconnaître pour une durée d'une seconde de parole choisie un locuteur parmi 630 avec moins de 5% taux d'erreur.

RÉFÉRENCES

[1] C. BARRAS, M.-J. CARATY, C. MONTACIE, "HMM based Acoustic-Phonetic Decoding with Constrained Transitions and Speaker Topology", Nato-ASI, 1993.

[2] C. MONTACIE, P. DELEGLISE, F. BIMBOT, M.-J. CARATY, "Cinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction", IEEE-ICASSP, San Francisco, 1992.

[3] C. MONTACIE, J.-L. LE FLOCH, "AR-Vector Models for Free-Text Recognition", ICSLP, pp. 475-478, 1992.

[4] C. MONTACIE, J.-L. LE FLOCH, "Discriminant AR-Vector Models for Free-Text Speaker Verification", Eurospeech 93, vol. 1, pp. 161-164, 1993.

[5] W. FISHER, V. ZUE, J. BERNSTEIN, D. PALLET, "An Acoustic-Phonetic Data Base" J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.

[6] J. J. WEBB, E. L. RISSANEN, "Speaker Identification Experiments Using HMMs", pp. 387-390, IEEE-ICASSP, 1993.

[7] B. S. ATAL, "Automatic Recognition of Speakers from Their Voices", IEEE Proceedings, vol. 64, No. 4, pp. 460-475, 1976.

[8] M. R. SAMBUR, "Selection of Acoustic Features for Speaker Identification", IEEE-TASSP, ASSP-23, pp. 176-182, 1975.

[9] A. E. ROSENBERG, F. SONG, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Dependent Modes", Computer Speech & Language, vol. 2, n. 3/4, pp. 143, 1988.

APPENDICE

Calcul des matrices de corrélation R_i :

$$R_i = \sum_{n=0}^{N-i-1} y_{n+i}\, y_n^t$$

Calcul des matrices A_i du modèle AR (algorithme Levinson-Whittle-Robinson) :

$A1_0^0 = I_d$; $A2_0^0 = I_d$; $D1_0 = R_0$; $D2_0 = R_0$;

for i = 0 to q-1 do

$$F_i = \sum_{j=0}^{i} A1_j^i R_{i+1-j}$$

$$K1_i = -F_i\, D2_i^{-1};\quad K2_i = -F_i^t\, D1_i^{-1}$$

$$D1_{i+1} = (I_d - K1_i K2_i)\, D1_i$$

$$D2_{i+1} = (I_d - K2_i K1_i)\, D2_i$$

$$A1_0^{i+1} = A1_0^i;\quad A2_0^{i+1} = K2_i\, A1_0^i$$

$$A1_{i+1}^{i+1} = K1_i\, A2_i^i;\quad A2_{i+1}^{i+1} = A2_i^i$$

for j = 1 to i do

$$A1_j^{i+1} = A1_j^i + K1_i\, A2_{j-1}^i$$

$$A2_j^{i+1} = A2_{j-1}^i + K2_i\, A1_j^i$$

end for

end for

for i = 0 to q do $A_i = A2_i^q$

Calcul du résiduel matriciel D :

$$X = \sum_{i=1}^{q}\left(\left(\left(\sum_{j=0}^{i} A_{i-j}R_i\right) + R_i\right)A_i^t\right)$$

$$Y = R_0 + \sum_{i=1}^{q}\left(A_i R_0 A_i^t\right)$$

$$D = X + X^t + Y$$

AVANTAGES D'UNE APPROCHE ANALYTIQUE ORIENTÉE CONNAISSANCES ENRECONNAISSANCE DU LOCUTEUR

Jean-François BONASTRE, Henri MÉLONI

LIUAPV, AVIGNON, FRANCE

Résumé

Cet article propose une approche orientée connaissances dans le cadre de la reconnaissance du locuteur utilisée en particulier pour l'étude explicite des paramètres spectraux spécifiques du locuteur.

Mots-clés :Identification, Spectral, Voix, Connaissances, Analytique.

1.INTRODUCTION

Nous nous intéressons à l'information spécifique du locuteur contenue dans les spectres à court terme. Dans ce cadre, nous étudions les capacités discriminantes de chaque phonème, associées à un ensemble d'outils de comparaison spectrale. Nous quantifions alors les diverses variations attachées aux représentations spectrales des phonèmes et nous diminuons leurs influences. Pour cela, nous utilisons une approche analytique orientée connaissances qui nous permet d'utiliser explicitement des connaissances expertes.

2.LA VARIABILITÉ DE LA PAROLE

La variabilité de la parole n'est pas uniforme mais composée de quatre types de variabilité : la variabilité aléatoire, la variabilité contextuelle, la variabilité interindividuelle et la variabilité de la langue [Rossi, 1989]. Dans un système informatique, nous rajoutons une cinquième cause de variabilité : la variabilité du traitement. Bien évidemment, toutes les causes de variabilité ne sont pas sources de problèmes pour la RAL, la variabilité interindividuelle est à la base de tout système de RAL. On cherche des indices maximisant cette variabilité, simples à mettre en œuvre, et le plus possible, indépendants du message (durée, composition grammaticale, etc.). La variabilité aléatoire, imprévisible dans sa forme mais quantifiable quant à ses effets, diminue la dynamique d'un système de calcul de ressemblance inter-locuteurs sans remettre en cause les indices utilisés. La variabilité de traitement conduit aux mêmes conclusions. Par contre, la variabilité contextuelle (et en particulier les effets de coarticulation) cause d'importants problèmes. L'apparence spectrale, l'emplacement des formants, la durée et l'accentuation d'un phonème sont modifiés par son contexte [Durand, 85]. Ces variations constituent une des grosses difficultés rencontrées dans les systèmes tolérant des énoncés de nature variable. Au contraire, la variabilité de la langue est généralement négligeable dans les applications pratiques.

3. COMPARAISONS SPECTRALES

La comparaison de formes spectrales est un problème difficile. Quelques points restent délicats, comme la prise en compte de la variabilité contextuelle, l'ajustement des niveaux d'énergie ou l'influence du mode de paramétrisation et de représentation des vecteurs spectraux. La réponse à ces problèmes n'est pas unique. Il faut en effet construire divers paramètres de comparaison, chacun adapté à la mise en évidence d'une information particulière.

Nous avons à notre disposition un ensemble de paramètres de comparaison permettant d'explorer toutes les caractéristiques des spectres à étudier [Bonastre, 94]. Chacune des distances spectrales définies comporte plusieurs possibilités d'adaptation : adaptation au type de représentation spectrale (Mel à 24 canaux ou linéaire à 128 canaux), au type de phonème (recherche de formants pour les voyelles, de pics d'énergie écartés dans le cas du /i/, ajustement énergétique ou non, etc.) ou à la dynamique recherchée (sélection des n plus mauvais canaux, utilisation d'un seuil majorateur, etc.). La comparaison entre les deux segments de signal s'effectue de manière optimale à l'aide d'une composition des diverses distances adaptées aux représentations spectrales (paramètre *COMP*, voir [Bonastre, 94]).

4.ÉVALUATION DES DIVERS PARAMÈTRES UTILISÉS

La variété de choix des indices discriminants amène une difficulté supplémentaire lors des phases de décision des systèmes : comment comparer un écart de x % dans la hauteur d'un formant avec un écart de durée d'un phonème ou de l'évolution de F0 ?

Historiquement la comparaison des indices est effectuée à l'aide des *F-ratio* [Wolf, 72 ; Calliope, 1989] :

$$F = \frac{\text{Variance des moyennes des différents locuteurs}}{\text{Moyenne de la variance intralocuteur}}$$

Soit en termes mathématiques :

$$F = \frac{\frac{1}{N}\sum_{i=1}^{i=N}\left(\overline{X} - \overline{X}_i\right)^2}{\frac{1}{N}\sum_{i=1}^{i=N}\frac{1}{M}\sum_{j=1}^{j=M}\left(\overline{X}_i - X_{ij}\right)^2}$$

avec :

- X_{ij} valeur de l'indice étudié pour le locuteur i et l'unité de test j.
- $\overline{X}_i$ moyenne de l'indice pour le locuteur i.
- $\overline{X}$ moyenne des moyennes des divers locuteurs.
- N nombre de locuteurs.
- M nombre de tests par locuteur.

Les *F-ratio* fournissent une évaluation des performances relatives des indices (a posteriori et à partir d'un grand nombre d'éléments analysés). Mais lors d'une décision à court terme–sur la foi de plusieurs indices et de quelques éléments d'analyse – les *F-ratio* ne permettent pas de mesurer la pertinence d'un indice par rapport à un autre et de déterminer si les informations obtenues suffisent à produire une décision. De plus, ce critère demande un grand nombre d'échantillons par individu pour garantir des résultats corrects (les moyennes sont calculées locuteur par locuteur, nécessitant pour chacun quelques dizaines d'échantillons).

D'autres critères d'évaluation peuvent apporter une solution, en considérant l'espace de variation intra et inter-locuteurs de chaque indice [Bonastre, 92 ; Van den Heuvel, 92 ; Thompson, 93]. Cette approche permet de prendre des décisions multi-indices avec une marge de sûreté configurable mais elle nécessite une étude préalable de chaque indice permettant la modélisation des espaces de variation. Les critères d'évaluation proposés accordent une grande importance au rapport de la variabilité inter-locuteurs sur la variabilité intra-locuteur sans tenir compte de l'écart des espaces de variations.

Pour essayer de remédier à ce problème, nous avons cherché une technique utilisant les moyennes globales inter et intra-locuteur ainsi que les écarts à la moyenne associés. Ce procédé doit accentuer l'influence des moyennes par rapport aux facteurs de dispersions [Bonastre, 94]. Nous proposons le critère de pertinence σm associant le rapport des écarts type intra et inter-locuteurs à la différence des moyennes inter et intra-locuteur. Ce qui nous donne la formule :

$$\sigma_m = \sqrt{\frac{\sigma_{\text{inter}}}{\sigma_{\text{intra}}}} . (m_{\text{inter}} - m_{\text{intra}})$$

avec

$$M_{inter} = \frac{\sum_{i=1}^{i=M} \left(\sum_{j=1}^{j=L_{Si}-1} D(S_i,R_j) + \sum_{j=L_{Si}+1}^{j=N} D(S_i,R_j) \right)}{M*(N-1)}$$

$$M_{intra} = \frac{\sum_{i=1}^{i=M} D(S_i,R_{L_{Si}})}{M}$$

$$\sigma_{inter} = \sqrt{\frac{\sum_{i=1}^{i=M} \left(\sum_{j=1}^{j=L_{Si}-1} (M_{inter} - D(S_i,R_j))^2 + \sum_{j=L_{Si}+1}^{j=N} (M_{inter} - D(S_i,R_j))^2 \right)}{M*(N-1)}}$$

$$\sigma_{intra} = \sqrt{\frac{\sum_{i=1}^{i=M} (M_{intra} - D(Si,R_{L_{Si}}))^2}{M}}$$

et :

- $D(Si,Rj)$ = Valeur de la ressemblance entre l'unité de test Si (qui peut être un spectre, un ensemble de spectre,ou une valeur de F0, ...) et la référence correspondante appartenant au locuteur j.
- L_{Si} correspond au locuteur qui a prononcé l'unité de test Si
- N correspond au nombre de locuteurs présents dans la base.
- M correspond au nombre d'unités de test présents dans la base.

NB : Toutes les moyennes et les écarts associés sont calculés sur l'ensemble des locuteurs et non locuteur par locuteur.

Pour avoir une vision des possibilités pratiques d'un système fondé sur ces techniques, nous calculons également le nombre de phonèmes nécessaires à l'identification d'un locuteur, selon le coefficient de sécurité choisi. Ce qui correspond au nombre d'occurrences garantissant la séparation des espaces de variation intra et inter-locuteurs. Nous recherchons le nombre d'occurrences nécessaires pour que l'intervalle de confiance à la moyenne intra-locuteur soit disjoint de l'intervalle de confiance à la moyenne inter-locuteurs. La formule correspondante est:

$$Nb_{occ} \geq (a^2 . \left(\frac{\sigma_{\text{intra}} + \sigma_{\text{inter}}}{M_{\text{inter}} - M_{\text{intra}}} \right)^2$$

(avec a fonction du coefficient de sécurité souhaité ; a≈2 pour un coefficient de 95%, a≈3 pour 99,7%)

Remarques :

- *Les valeurs obtenues définissent le cas le plus défavorable : utilisation de plusieurs occurrences d'un même phonème (redondance des informations), pour chaque occurrence on considère que l'imposteur est le locuteur le plus proche de l'individu, etc.*
- *Durant les tests de reconnaissance, une identification est dite réussie si l'indice prend pour le locuteur concerné une valeur strictement inférieure à celles des autres locuteurs de la base.*

5.INFORMATIONS SPECTRALES

Nous proposons de référencer les caractéristiques individuelles contenues dans les spectres à court terme et de quantifier les variations affectées aux représentations spectrales des divers phonèmes du français.

1.DESCRIPTION DE L'EXPÉRIENCE

Nous avons choisi un corpus extrait de la base BDSON, composé d'un énoncé étiqueté ("la bise et le soleil") d'environ 50 secondes prononcé par 22 locuteurs. Ce corpus est constitué d'un texte unique.

Dans le but de comparer différentes méthodes de paramétrisation du signal, nous avons calculé les spectres représentant les phonèmes de trois manières différentes : à l'aide d'une transformée de Fourier rapide, par une prédiction linéaire à 14 coefficients et à 21 coefficients. Les spectres sont modélisés de deux manières : par 24 valeurs réparties suivant une échelle de Mel et par 128 valeurs réparties linéairement.

Nous mesurons pour un spectre dans un contexte donné (et à l'aide du paramètre étudié) la distance à la référence correspondante du locuteur concerné (distance intra-locuteur). De la même manière, nous calculons les distances aux références des autres locuteurs de la base (distances inter-locuteurs). Ce travail est effectué pour chaque spectre caractérisant un segment phonétique de l'énoncé. Nous évaluons alors la distance moyenne d'un spectre à la référence correspondante du locuteur concerné (ainsi que l'écart à la moyenne). Le même type de calcul est effectué pour calculer la distance moyenne inter-locuteurs (moyenne des distances entre chaque spectre du signal et les références des autres locuteurs) et l'écart à la moyenne inter-locuteurs. Ces résultats sont produits pour chacun des paramètres de comparaison étudiés et pour chaque mode de paramétrisation du signal.

Comme il était difficile de séparer les corpus d'apprentissage et de reconnaissance, le premier spectre rencontré dans un contexte donné est stocké comme référence, les suivants sont mémorisés dans la base de reconnaissance.

2.INFLUENCE DU MODE DE PARAMÉTRISATION DU SIGNAL

Le passage d'un signal de parole brut à un spectrogramme peut être effectué de plusieurs manières. Une technique consiste à considérer ce signal de parole comme un signal quelconque et à le traiter de manière standard, par exemple à l'aide d'une transformée de Fourier. D'autres solutions s'inspirent des connaissances sur l'anatomie humaine et en particulier des organes et des stratégies intervenant dans la production de la parole.

Nous cherchons à savoir si une paramétrisation est réellement plus performante et si le gain obtenu en performance ou en agrément (facilité de visualisation des informations pour l'expert) justifie un coût supplémentaire éventuel en ressources calcul. Nous évaluons la pertinence de l'indice de comparaison *COMP* à l'aide du critère σm. Ces résultats sont estimés pour chaque fenêtre de 320 échantillons[1] à l'aide d'une analyse FFT sur 128 points, d'une modélisation LPC à 14 coefficients et d'une LPC à 21 coefficients. Les références sont acquises en fonction du contexte (une référence pour chaque triplet phonémique et chaque locuteur).

		Mode de Paramétrisation		
Phonème	Nb tests	FFT	LPC 14	LPC 21
Voyelles	1026	148	214	255
Fricatives	259	441	605	595
Occlusives	390	165	207	211
Liquides	356	133	170	183
Semi-V.	171	179	258	262
Total	**2202**	**185**	**257**	**264**

Table 1 : Tableau donnant pour chaque classe phonétique, le nombre de mesures intra-locuteur et inter-locuteurs et le pouvoir de discrimination de l'indice COMP, estimé à l'aide du critère σm, suivant le mode de paramétrisation utilisé.

Les résultats, décrits dans la table 1, montrent l'influence d'un modèle de production comme le modèle LPC sur des indices de comparaison puisque le gain moyen est supérieur à 30%. Par contre, l'amélioration obtenue en passant d'un modèle à 14 coefficients à un modèle à 21 coefficients (au mieux de 10%) ne justifie que peu l'accroissement des calculs.

Remarques :

- *Il faut noter que les paramètres de comparaison sont mis au point à partir d'observations expertes elles-mêmes réalisées à l'aide de spectrogrammes de type LPC.*

- *Il est difficile de dissocier les modes de représentation spectrale des indices de comparaison, les paramètres étant définis spécifiquement pour une représentation donnée ou, tout au moins, optimisés pour un type de vecteurs spectraux.*

3.INFLUENCE DU CONTEXTE

Le problème principal rencontré par des systèmes de RAL, utilisant comme critère sélectif des comparaisons spectrales, concerne les phénomènes de coarticulation. Ceux-ci amènent une grande variabilité spectrale des phonèmes. Cette variabilité, engendrée par la coarticulation, est en relation directe avec le contexte des phonèmes. Il est nécessaire, dans le cadre d'un système de RAL, de mesurer et de diminuer au maximum l'influence de ces phénomènes.

Nous avons donc cherché dans quelle mesure le contexte immédiat gauche et droit influence l'aspect spectral d'un phonème (l'influence est ressentie à plus long terme, mais il n'est pas envisageable, pour des raisons évidentes de complexité, de travailler sur des n-uplets phonétiques supérieurs à des triplets).

Nous calculons la sélectivité des paramètres phonémiques à l'aide du critère σm, du pourcentage d'identification réussie pour une occurrence et du nombre d'occurrences nécessaires pour séparer les espaces de variations intra-locuteur et inter-locuteurs avec un facteur de sécurité de 95%.

[1] Une fenêtre représente 1/100ème de seconde. 100 vecteurs sont donc calculés par seconde avec un recouvrement des fenêtres.

Ces résultats sont recherchés avec différentes méthodes de prise en compte du contexte phonétique immédiat :
- utilisation du contexte gauche et droit : une référence est mémorisée pour chaque triplet phonétique,
- une seule référence par phonème : connaissance du phonème mais pas du contexte,
- utilisation du contexte gauche : une référence correspond à un doublet phonétique composé du phonème précédent et du phonème étudié,
- utilisation du contexte droit : une référence par doublet.

On peut remarquer, table 2, l'influence de la prise en compte du contexte sur les résultats, que ce soient en terme d'identifications réussies (de 20% à 49%) ou en nombre d'occurrences nécessaires pour obtenir une marge de sécurité de 95% (de 60 à 10). Les capacités discriminantes des phonèmes–mesurées à l'aide de l'indice σm–varient aussi de près de 50% suivant la prise en compte du contexte.

Phonème	tests	σm	id%	Nb_{occ}
une référence par triplet phonétique				
Fricatives	259	596	57%	7
Occlusives	390	211	47%	12
Liquides	356	184	38%	14
Semi-voy.	171	263	57%	6
Voyelles	1026	226	50%	8
Total	2202	265	49%	10
une référence par phonème				
Fricatives	1081	410	31%	14
Occlusives	1265	100	21%	56
Liquides	1304	40	14%	430
Semi-voy.	286	102	30%	100
Voyelles	3579	81	18%	62
Total	7515	122	20%	60
prise en compte du contexte gauche (doublet)				
Fricatives	538	416	38%	14
Occlusives	672	118	30%	38
Liquides	794	114	25%	52
Semi-voy.	218	184	50%	20
Voyelles	2054	132	35%	25
Total	4276	168	34%	30
prise en compte du contexte droit (doublet)				
Fricatives	577	542	47%	9
Occlusives	611	182	39%	16
Liquides	917	136	28%	24
Semi-voy.	221	240	53%	7
Voyelles	1980	160	37%	16
Total	4306	220	38%	17

Table 2 : Pour chaque classe phonétique, à l'aide de *COMP* et suivant le mode de prise en compte du contexte : nombre de tests réalisés, σm, pourcentage d'identifications réussies à partir d'une occurrence, nombre N d'occurrences nécessaires pour obtenir un facteur de sécurité de 95%.

Quelques résultats et commentaires :
- L'utilisation d'une référence par triplet offre des résultats très nettement supérieurs aux autres propositions, conformément à notre attente.
- La comparaison des résultats obtenus en prenant en compte les contextes gauche et droit tourne à l'avantage du contexte droit. Comme on pouvait s'y attendre, l'anticipation d'un son influe plus sur l'apparence acoustique du phonème étudié que les contraintes articulatoires dues au phonème précédent, ce qui accentue la variabilité intra-locuteur et diminue l'efficacité des indices utilisés.
- L'influence du contexte, lors de l'utilisation d'une seule référence par phonème, est particulièrement importante sur les liquides et les semi-voyelles. Les paramètres basés sur l'étude des voyelles sont également atteints par la variabilité contextuelle (65% de diminution du critère σm) mais plus légèrement. Les fricatives semblent moins marquées par la coarticulation. Les occlusives restent dans la moyenne, sensibles à la variation contextuelle mais conservant des résultats acceptables.

6. COMPARAISON AVEC UNE MÉTHODE "AVEUGLE"

Les choix technologiques que nous avons effectués privilégient l'approche orientée connaissances et nous permettent d'utiliser, au sein des paramètres de comparaisons de formes spectrales, la connaissance explicite du phonème et de son contexte. Cette approche, malgré ses avantages– évoqués dans le paragraphe précédent–présente l'inconvénient de nécessiter une lourde phase de localisation des phonèmes. Nous voulons comparer nos choix avec une technique classique, proche de la quantification vectorielle, de manière à montrer que les avantages de l'approche AOC justifient les difficultés de localisation.

Dans cet esprit, nous avons conservé le corpus précédent, extrait de la base BDSONS. Nous effectuons également l'acquisition des références de la même manière que précédemment, soit une référence par triplet phonétique et par locuteur, obtenue à partir de l'étiquetage standard fourni avec le corpus. Nous conservons le mode de paramétrisation du signal (LPC-21) et la méthodologie générale d'extraction d'informations statistiques.

Par contre, l'évaluation de la ressemblance entre une unité du signal et les références des divers locuteurs est effectuée en aveugle. Pour un locuteur, la note attribuée à l'unité est calculée en comparant l'unité du signal à chaque référence appartenant au locuteur, le résultat correspond alors à la distance minimum trouvée pour ce locuteur.

Les conditions de l'expérience se résument en quelques remarques :
- La paramétrisation du signal et le mode de représentation des spectres restent inchangés (LPC-21 associé à une échelle de Mel sur 24 canaux et une échelle linéaire à 128 canaux).
- L'apprentissage est réalisé avec un étiquetage manuel explicite (une référence par locuteur et par triplet phonétique distinct).

- Une unité de test est définie par l'étiquetage manuel (sélection d'une trame représentant l'unité à l'aide de l'étiquetage) mais sans conserver d'information sur sa nature (ni connaissance du phonème, ni de son contexte).
- La note attribuée à l'unité de test, pour un locuteur donné, correspond à la distance minimale obtenue en comparant l'unité de test à chaque référence appartenant à ce locuteur.
- Le paramètre de comparaison spectrale *COMP* est optimisé suivant le phonème en contexte étudié. Dans le cadre de cette expérience, nous prenons en compte la connaissance de la référence et du contexte dans lequel elle a été acquise, en ignorant la nature de l'unité étudiée.

	comparaisons AOC				comparaisons "aveugle"			
pho	id%	σ_l/σ_i	m_{in}-m_{loc}	Nb occ	id%	σ_l/σ_i	m_{in}-m_{loc}	Nb occ
f	42%	46%	397	7	26%	98%	63	16
s	59%	30%	576	4	29%	111%	70	18
v	50%	63%	261	4	31%	85%	64	19
z	64%	106%	209	7	39%	111%	82	11
p	29%	99%	143	18	18%	93%	38	66
t	41%	82%	215	11	20%	99%	45	42
k	33%	119%	136	26	21%	104%	48	40
b	58%	64%	255	5	23%	95%	64	31
d	53%	68%	266	6	30%	103%	63	26
m	62%	81%	199	5	44%	110%	94	13
l	34%	92%	171	15	26%	103%	55	30
r	46%	79%	176	13	25%	106%	54	36
j	84%	71%	302	2	41%	103%	85	10
a	54%	71%	221	6	28%	104%	63	27
ɔ	50%	102%	161	8	24%	99%	60	29
ɛ	60%	66%	218	3	23%	100%	56	27
o	70%	95%	159	11	16%	101%	50	52
e	51%	68%	291	6	24%	96%	57	30
ə	29%	63%	146	14	19%	88%	51	35
i	48%	97%	192	6	21%	100%	50	32
y	43%	41%	258	6	15%	98%	36	47
u	48%	59%	228	9	20%	103%	44	54
ã	62%	52%	182	5	21%	88%	56	33
õ	61%	86%	177	4	22%	101%	52	33
Tot	49%	67%	216	10	26%	101%	60	27

Table 3 : Comparaison des résultats d'identification obtenus en connaissant le phonème et son contexte et en mode aveugle (sauf pour l'apprentissage) ; pourcentage d'identifications réussies à l'aide d'une seule occurrence, rapport des écarts type inter/intra-locuteur, différences entre les moyennes inter et intra-locuteur et N, nombre de phonèmes nécessaires à une identification avec un facteur de sécurité de 95% (le tout à l'aide de *COMP*).

La comparaison des résultats obtenus par une méthode utilisant la connaissance explicite du phonème et de son contexte avec les scores obtenus par une méthode de type "aveugle" tourne nettement à l'avantage de l'approche AOC (Table 3). Le taux d'erreur est réduit d'un tiers alors que la durée nécessaire pour garantir une marge de sécurité de 95% est divisée par trois.

Les phonèmes semblent tous réagir de même manière : variations inter et intra-locuteur de même ampleur et normalisation des scores des distances inter-locuteurs et intra-locuteur (réduction très nette de la différence entre les moyennes inter et intra-locuteur).

7.CONCLUSION

Nous avons tout d'abord mis en évidence l'information spécifique du locuteur véhiculée par les spectres à court terme représentatif des sons du français.

L'influence du contexte environnant sur l'apparence spectrale d'un phonème, nous est apparue génératrice de variations masquant la spécificité du locuteur. Après en avoir mesuré les effets, nous nous sommes attachés à définir des techniques minimisant l'importance de ce phénomène en prenant en compte de façon explicite le phonème étudié et son contexte. Ces travaux nous ont permis de démontrer qu'un faible nombre de phonèmes suffisait pour rechercher, au sein d'un ensemble de personnes connues, le locuteur présentant la plus forte ressemblance avec un individu caractérisé par un message vocal, et ce avec un facteur de sécurité très important. Cependant, notre méthodologie implique une participation volontaire des locuteurs et un environnement sensiblement identique lors de l'apprentissage et des tests de reconnaissance.

En comparant une approche utilisant explicitement une partie des informations symboliques disponibles avec des techniques de type "aveugle", nous avons montré que les méthodes analytiques orientées connaissances obtenaient des performances très nettement supérieures (pour une marge de sécurité comparable, la durée nécessaire à une identification est divisée par trois) ce qui justifie pleinement le coût d'une localisation des éléments d'analyse dans le cadre d'un système d'identification automatique du locuteur.

A partir de l'approche présentée ici, nous avons réalisé un démonstrateur [Bonastre, 93, 94] qui obtient des résultats comparables aux méthodes algébriques (cependant, il est nécessaire de fournir au système la composition phonémique théorique de l'énoncé).

Nous souhaitons rentabiliser l'approche Analytique Orientée Connaissances que nous avons utilisée en intégrant au système des informations de nature différente (prosodiques, micro-prosodiques, étude des transitions, etc.) ce qui devrait, outre l'étude des capacités discriminantes de ce type d'informations, améliorer les performances obtenues.

Pour réaliser cela, nous envisageons d'utiliser conjointement à notre technique d'autres approches, comme les modèles de Markov ou les modèles Auto-Régressifs Vectoriels qui semblent bien adaptées à l'étude de phénomènes dynamiques. Dans ce cas, les connaissances expertes superviseraient les approches de type "boite noire".

En outre, dans le cadre de ce travail, nous avons senti l'intérêt qu'il y aurait à évaluer l'influence des facteurs temporels (vieillissement de la voix) et psychophysiologiques sur la variabilité et la spécificité du locuteur. Ces divers points feront l'objet d'une prochaine série de travaux.

BIBLIOGRAPHIE

[Bonastre 92] J.F. Bonastre, H. Méloni, *Etude de la variabilité spectrale pour la caractérisation du locuteur* ; 19èmes JEP, Bruxelles, Belgique.

[Bonastre 93] J.F. Bonastre, H. Méloni ; *Speaker Recognition and Analytic Process* ; Proc. 3rd European Conference on Speech Communication and Technology, 21-24 septembre, Berlin, Germany.

[Bonastre 94] J.F. Bonastre ; *Stratégie Analytique Orientée Connaissances pour la Caractérisation et l'Identification du Locuteur ;* Thèse de l'université d'Avignon.

[Calliope 89] Calliope ; *La parole et son traitement automatique* ; Masson.

[Durand 85] P. Durand ; *Variabilité acoustique et invariance en Français consonnes occlusives et voyelles* ; Edition du CNRS.

[Rossi 89] M. Rossi ; *De la quiddité des variables* ; Actes du Séminaire Variabilité et spécificité du locuteur, Marseille Luminy, 20-21 juin 1989, pp. 11-31.

[Thompson 93] J. Thompson, J. S. Mason ; *Within class optimization of cepstra for speaker recognition* ; Proc. 3rd European Conference on Speech Communication and Technology, 21-24 septembre, Berlin, Germany, pp. 165-168.

[Van den H. 93] H. Van den Heuvel, T. Reitveld ; *Speaker Related Variability in Cepstral Representations of Dutch Speech Segments*; ICSLP 92, Canada, Vol. 2, pp.1581-1584.

[Wolf 72] J. Wolf ; *Efficient acoustic parameters for speaker recognition* ; JASA, Vol. 51, pp. 2045-2056

Recherches comparatives pour caractériser des locuteurs, en particulier des jumeaux

M. Deppe, W. Endres, M. Fischer

Technische Hochschule de Darmstadt, Institut de Communication et d'Electroacoustique

Résumé

Le rapport donné traite de recherches sur la ressemblance des voix de jumeaux homozygotes et de frères normaux. Pour trois voyelles plusieurs caractéristiques de la voix sont étudiées. A l'aide de tests significatifs deux mesures de ressemblance ont été développées. Par comparaison on trouve, que les valeurs de ressemblance sont sensiblement supérieures en cas des jumeaux homozygotes.

Mots-clés: Reconnaissance du locuteur, voix des jumeaux homozygotes et des frères, mesures de ressemblance

1. INTRODUCTION

Nos études avaient le but de comparer les voix des homozygotes et des frères (ou soeurs) hétérozygotes en ce qui concerne leur ressemblance. Des connaissances à ce sujet tout intéressantes à la fois du point de vue anthropologique et technique. Pour l'anthropologie, il s'agit de savoir si les caractéristiques de voix sont déterminées génétiquement. Une distinction des homozygotes pourrait être très importante en cas de reconnaissance du locuteur.

2. MANIERE D'AGIR

Il y avait cinq paires d'homozygotes et cinq paires de frères hétérozygotes comme sujets. Ici il était nécessaire de se borner aux sujets masculins, la fréquence fondamentale de la voix étant plus basse par rapport à la fréquence fondamentale de la voix des sujets féminins et le nombre des harmoniques est donc plus grand pour une bande de la fréquence donnée. Pour réduire l'influence occasionnée par des maladies lors de la puberté et pour diminuer la probabilité d'un changement du système d'articulation par suite d'une opération ou d'une maladie, nous avons seulement fait des analyses avec des personnes entre 18 et 30 ans. Les trois premières fréquences moyennes des formants de la voyelle /a/,les deux premières fréquences moyennes des formants des voyelles /e/ et /i/ et de même les fréquences fondamentales respectives étaient utilisées comme caractéristiques comparatives. Il s'agit en somme de dix caractéristiques. En outre nous avons étudié les modifications temporelles de ces paramètres au cours des phrases. Pour assurer un milieu naturel aussi bon que possible des voyelles et pour avoir la possibilité d'épuiser toute la gamme de variations, il était nécessaire de construire des phrases convenables. Au total 13 variantes différentes de la voyelle /a/ et 15 ou 12 variantes différentes des voyelles /e/ et /i/ ont été étudiées. Pour l'enregistrement chaque locuteur devait prononcer deux fois les phrases d'essai, donc il existait deux corpora complets. Les enregistrements ont été faits avec un DAT-recorder dans un laboratoire avec isolation acoustique. Les phrases déjà mentionnées et enregistrées en fonction du temps étaient seulement intéressantes par rapport aux voyelles. Ces fonctions ont été divisées en tranches de 100 ms, dans lesquelles on peut considérer le signal de la parole comme une fonction quasi-périodique. Après avoir amplifié l'énergie des hautes fréquences, chaque tranche était soumise à une 1024 points-FFT en employant une fenêtre de Hamming correspondant à une décomposition en fréquences de 11,72 Hz. Utilisant ces spectres il est possible de trouver exactement les fréquences du centre des formants et les fréquences fondamentales de la parole. Toutes ces données ont servi de base pour les évaluations statistiques et aussi pour le calcul des taux de ressemblance.

3. ETUDE DE L'AFFINITE

Pour exécuter des tests de l'affinité entre les sujets il y avait quatre possibilités de combinaison:

- les deux échantillons de parole d'un seul locuteur (auto-test)
- des échantillons de parole des homozygotes
- des échantillons de parole des frères hétérozygotes
- des échantillons de parole de personnes quelconques, qui n'étaient pas parentes.

Dans les trois derniers cas, nous avons respectivement réunis les échantillons de parole d'un locuteur en un échantillon. A l'aide de l'auto-test il était ainsi possible de juger la qualité de cette méthode. Naturellement les locuteurs identiques doivent avoir la similarité la plus grande. Au contraire la comparaison des sujets quelconques était utilisée comme une délimitation contre des ressemblances aléatoires. Puis la comparaison des limites d'affinité définies dans ce qui suit obtenues des échantillons de parole des homozygotes et des frères hétérozygotes était très importante.

La mesure de ressemblance était définie comme jauge d'affinité pour la caractéristique des fréquences moyennes des formants. Cette définition est fondée sur des comparaisons statistiques, dans lesquelles les données d'un échantillon de parole sont interprêtées comme un échantillonnage d'un ensemble fondamental. Ici nous avons employé le double test de l'échantillonnage pour la différence de deux valeurs moyennes arithmétiques avec une probabilité de 99 %. Pour l'autotest nous avions pris pour base des échantillonnages conjointes, pour toutes les autres comparaisons des non-conjointes. A cette occasion nous sommes partis de l'hypothèse, que les deux échantillons de parole étaient ceux d'un même locuteur. Les tests de signification ont été calculés individuellement pour chacune des dix caractéristiques susmentionnées. Le résultat de chaque test pouvait être le rejet ou l'acceptation de cette hypothèse. Si l'on attribue un 0 à chaque rejet et un 1 à chaque acceptation, on peut trouver par accumulation un indice pour la similitude - la valeur de ressemblance. En conséquence la mesure de ressemblance apparaît seulement en nombres cardinaux entre 0 et 10, et on peut déduire que plus grande est cette valeur, plus grande est la ressemblance.

Pour la caractéristique dynamique des évolutions temporelles des fréquences des formants ou des contours, nous avons défini pour chaque formant la valeur euclidienne de ressemblance sur la base des distances euclidiennes entre les fréquences de voyelles attribuées chaque fois aux cours des phrases. Pour chacun des dix cours des contours il était possible de trouver une valeur caractéristique par le cumul de la racine carrée de toutes les distances. Si on additionne ces dix valeurs caractéristiques on peut utiliser cette somme pour définir un nouvel indice de ressemblance. En effet il a fallu ici tenir compte que le nombre des voyelles considérées différait. En plus la largeur des formants s'agrandit si la fréquence s'élargit; cela veut dire que les valeurs des distances euclidiennes s'élargissaient aussi. Pour éviter que les formants élevés ne reçoivent une importance trop forte nous nous avons rapportés pour chaque formant aux fréquences moyennes des formants des deux échantillons de parole comparés l'un avec l'autre.

Pour l'indice d'affinité IA il s'ensuit

$$IA = \frac{\sqrt{\sum_{i=1}^{m(k)} (x_{1ki} - x_{2ki})^2}}{\frac{n(k) \cdot (\overline{X}_{1k} + \overline{X}_{2k})}{2}} \cdot q$$

q = constant

k = nombre des contours examinés
m(k) = nombre des voyelles examinées à l'intérieur d'une phrase
n(k) = fréquences des formants comme valeurs moyennes, prises aux échantillons des phrases

Si les constantes q sont choisies convenablement l'affinité euclidienne a aussi des valeurs entre 0 et 10. En ce cas on doit en effet interpréter les résultats en sens inverse aux valeurs de la jauge de ressemblance. Cela veut dire que si les valeurs sont basses, une grande affinité entre les locuteurs est possible, si les valeurs sont très hautes, les différences peuvent être significatives.

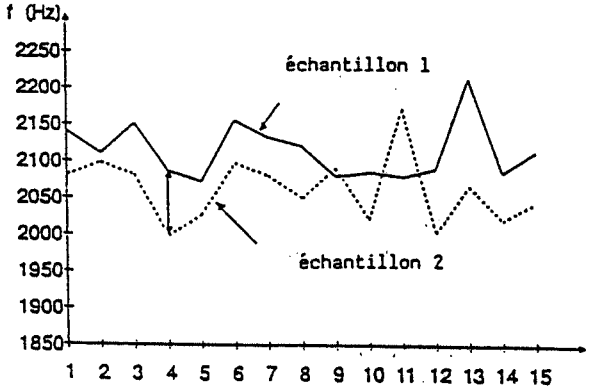


Fig. 1 Voyelles attribuées pour la recherche de la distance euclidienne (2ème formant de /e/).

4. RESULTATS

Des échantillons de parole qui étaient à notre disposition, il était possible de déduire 210 combinaisons de paires que l'on pouvait diviser en quatre groupes comme déjà mentionné. Les résultats sont démontrés dans ce qui suit. Nous avons commencé par la présentation de la jauge de ressemblance, cela veut dire, avec une comparaison des fréquences moyennes des formants.

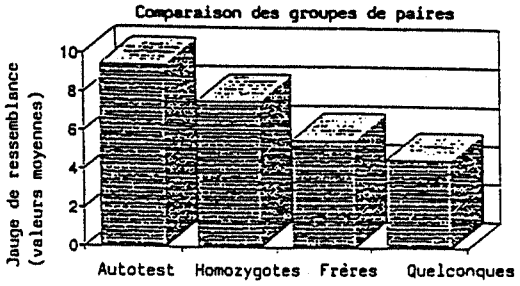


Fig. 2: Valeurs en moyennes des jauges de ressemblance pour quatre groupes d'une paire

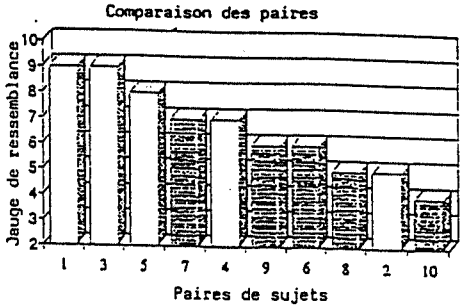


Fig. 3: Comparaison des valeurs de la jauge de ressemblance des monozygotes (blanc) et des paires dizygotes de frères par ordre de gradeur

La comparaison des échantillons de parole d'un seul et même locuteur a démontré que ce dernier présentait absolument les plus hautes valeurs des indices de ressemblance (fig. 2). Par ce résultat l'aptitude de cette méthode pouvait être démontrée, parce que naturellement l'indice de ressemblance entre des locuteurs identiques doit être la plus haute. En comparaison, les valeurs des indices de ressemblance des sujets combinés de toute manière étaient évidemment plus basses en moyenne que les valeurs trouvées pour les homozygotes, les hétérozygotes et lors de l'autotest. Donc on peut admettre, que la similarité des paires de frères n'existe pas par hasard. Au contraire cela est plutôt dû à des influences héréditaires et à l'environnement. Si l'on compare les paires un à un (Fig. 3) on trouve, qu'en effet il n'est pas possible de séparer les paires homozygotes nettement des paires hétérozygotes, mais que la plupart des homozygotes ont cependant des valeurs plus hautes d'indices de ressemblance. On pouvait constater, que l'indice de ressemblance moyenne de l'ordre de 7,6 en cas des homozygotes était sensiblement plus haute que la valeur 5,6 en cas des hétérozygotes. Faisant usage de l'indice de ressemblance euclidienne on trouvait des valeurs comparables.

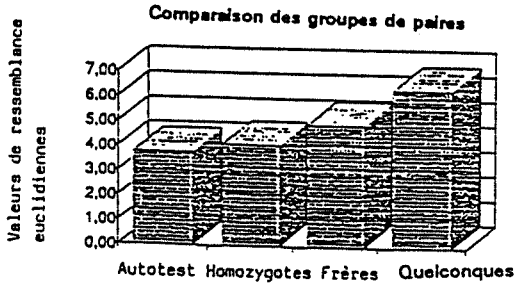


Fig. 4: Valeurs de ressemblance euclidienne pour les quatre groupes de paires

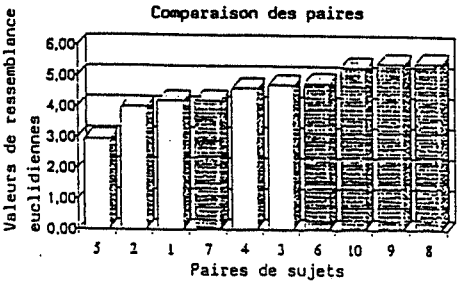


Fig. 5: Comparaison des valeurs de ressemblance euclidienne des jumeaux homozygotes (blanc) et des paires dizygotes masculins, classées par ordre de grandeur

Ici aussi, nous pouvons confirmer l'utilité de ces méthodes avec un autotest. Cela veut dire, on peut constater une nette délimitation par rapport aux valeurs de ressemblance par hasard (Fig. 4). En comparant les paires individuellement (Fig. 5) on peut constater évidemment un indice de ressemblance plus haute des homozygotes aussi en utilisant ces valeurs.

En conséquence on peut constater qu'une concordance des voix des homozygotes est signficativement plus haute que celle des paires de frères hétérozygotes en employant non seulement l'indice de ressemblance à base des fréquences moyennes des formants mais aussi à base des modifications temporelles des contours des fréquences moyennes des formants.

En utilisant cette méthode des homozygotes on peut maintenant estimer, dans quelle mesure ces symptômes peuvent être considérés comme héréditaires ou si des influences ambiantes sont prédominantes. Ici nous avons profité du fait que des homozygotes sont complètement identiques. Cela veut dire, que des différences de prononciation se ramènent seulement

aux influences ambiantes, pendant que, en cas des paires des frères hétérozygotes non seulement des différences génétiques mais aussi ambiantes peuvent exister. Si l'on suppose que les différences à l'intérieur de la parole se présentent dans nos recherches des quotes-parts de la ressemblance comme des rejets et qu'en cas des frères monozygotes et hétérozygotes les différences des influences ambiantes ne sont en effet pas absolument de même nature mais environ de même intensité, on peut supposer:

$$\text{rejet}_{\text{homozygotes}} = \text{influences ambiantes}_{\text{homozygotes}}$$

$$\text{rejet}_{\text{frères}} = \text{génétique} + \text{influences ambiantes}_{\text{frères}}$$

$$\text{influence ambiante}_{\text{homozygotes}} \approx \text{influence ambiante}_{\text{frères}}$$

En utilisant les résultats de nos recherches on peut en conséquence définir le quotient suivant:

$$\frac{\text{influence ambiante}}{\text{génétique}} = \frac{EA_{\text{homozygotes}}}{EA_{\text{frères}} - \text{rejet}_{\text{homozygotes}}}$$

$$= \frac{2,4}{4,4 - 2,4} = 1,2$$

Pour la mesure euclidienne on peut supposer que ces valeurs représentent immédiatement les différences dans la parole:

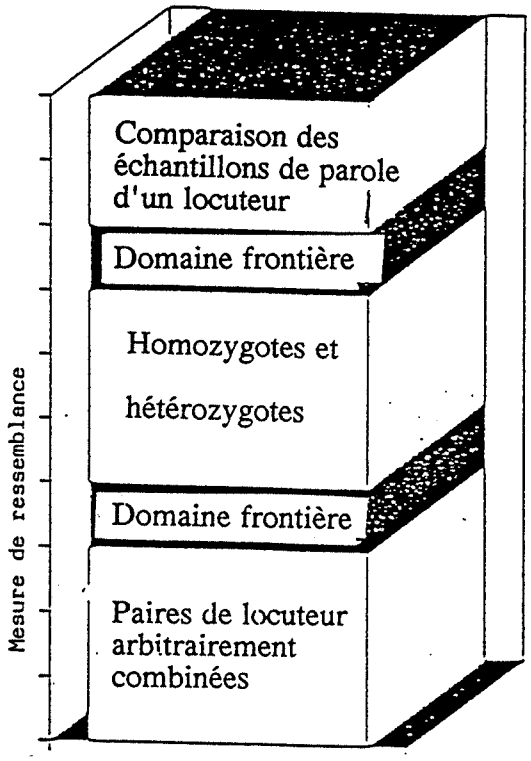
$$\frac{\text{influence ambiante}}{\text{génétique}} = \frac{EA_{\text{homozygotes}}}{EA_{\text{frères}} - EA_{\text{homozygotes}}}$$

$$= \frac{4,09}{4,93-4,09} = 4,87$$

Donc on peut constater que les fréquences moyennes des formants sont influencées à parts égales par la génétique et par des influences ambiantes, tandis que leur déroulement dans le temps, c'est à dire la prosodie, représentée par la mesure de ressemblance euclidienne, dépend évidemment davantage des influences ambiantes.

5. CONCLUSION

Même si, pour pouvoir étendre ces résultats, il faut un plus grand nombre d'échantillons, on peut constater que la présente étude tend à montrer que la ressemblance des jumeaux homozygotes est plus grande que celle des frères hétérozygotes, cela aussi bien pour la prosodie soumise à une forte influence ambiante que pour les fréquences de milieu moyennes des formants à base plutôt morphologique. Cela pose des exigences plus élevées pour la capacité de différenciation des systèmes de vérification du locuteur.



BIBLIOGRAPHIE

Alpert, M.; Kurtzberg, R.L.; Pilot, M. & Friedhoff, A.J. (1963)
Comparison of the Spectra of the Voices of Twins.
J. Acoust. Soc. Am. 35, 1877 (A)

Gedda, L.; Fiori-Ratti, L. & Bruno, G. (1960)
La voix chez les jumeaux monozygotiques.
Phoniatrica 12, 81-94

Kersta, L.G. & Lolangelo, J.A. (1970)
The Spectrographic Speech Patterns of Identical Twins.
J. Acoust. Soc. Am. 47, 56-59 (A)

Künzel, H.J.
Sprecherkennung, Grundzüge forensischer Sprachverarbeitung
Kriminaltechnik Verlag, Heidelberg 1987

ETUDE DES VARIABILITÉS INTRA- ET INTER-LOCUTEURS PROVOQUÉES PAR L'ÉMOTION ET L'IMITATION POUR LES SYSTÈMES DE VÉRIFICATION DU LOCUTEUR

M. Mehdi Homayounpour (*), Gérard Chollet (+)(**), Jacqueline Vaissière (*)

(*) CNRS/URA 1027, 19 rue des Bernardins, 75005 Paris, France
(+) IDIAP, C.P. 609 , 1920, Martigny, Suisse
(**) TELECOM-Paris, ENST, Dept. Signal, CNRS-URA 820, 46 rue Barrault, 75634 Paris, France

Résumé:

Les systèmes actuels de vérification automatique du locuteur (SVAL) obtiennent de bonnes performances dans les cas où les locuteurs parlent d'une voix relativement neutre, c'est-à-dire sans exprimer une émotion particulière, et dans les cas où ils n'essaient pas d'imiter une autre voix. Un SVAL a été élaboré, basé sur des coefficients LPC cepstraux (LPCC) et sur l'algorithme de programmation dynamique (DTW). Une méthode classique d'évaluation d'un SVAL est de comparer sa performance à celle obtenue par des auditeurs humains. Nous avons comparé les performances de notre propre système avec celles obtenues par des auditeurs humains sur des données présentant quatre états émotifs différents (neutre, joie, fatigue et colère).

Nous avons constaté que les variations des contours de F0, de la vitesse d'élocution, de l'intensité, des valeurs des formants, etc... dues à l'émotion exprimée dégradent les performances de notre système automatique et des auditeurs humains. Dans une étude spectrographique, nous avons étudié les variations prosodiques et spectrales provoquées par l'émotion. Dans une autre expérience, une tâche de vérification du locuteur a été effectuée automatiquement et par des locuteurs humains sur une base de données de parole imitée, enregistrée au téléphone: 24 locuteurs essayaient d'imiter deux locuteurs de référence. Le système automatique et les auditeurs humains ont des performances très proches.

Mots-clés: vérification du locuteur, émotion, imitation

1. INTRODUCTION

De nombreux systèmes ont été élaborés et différentes méthodes ont été proposées pour vérifier l'identité proclamée par un locuteur. Ces systèmes ont généralement obtenu une bonne performance dans des conditions contrôlées. Des études séparées [1] [2] ont mis en évidence des variations de la parole dues à l'état émotionnel des locuteurs. Il était important de tester dans quelle mesure ces variations pouvaient affecter les performances des SVAL et de disposer de données sur la spécificité acoustique d'énoncés émotionnels. Une recherche systématique des changements des paramètres acoustiques de la parole liée aux états émotifs de 12 locuteurs a été réalisée et quelques expériences ont été conduites pour déterminer les effets de l'émotion dans les tâches de reconnaissance du locuteur. Ces effets ont également été étudidés dans [4] [5].

2. BASES DE DONNÉES

Deux bases de données ont été enregistrées, l'une pour étudier les effets de l'émotion et l'autre les effets de l'imitation. Pour la première (Base I), la base de données de parole émotionnelle, 12 locuteurs français âgés de 24 à 55 ans et présélectionnés (4 locuteurs sur 16 éliminés) ont simulé successivement un état neutre, deux émotions (joie et colère) et un état physique particulier (fatigue). Pour la seconde base (Base II), la base de données de parole d'imitation, les locuteurs, hommes et femmes, ont tenté d'imiter une voix de leur sexe, parmi deux voix de référence, à travers le téléphone.

Après une phase d'entraînement (car il n'est pas évident de simuler différents états émotifs devant un micro), pour la première base, chaque locuteur a été enregistré dans une chambre sourde en deux sessions:
a) Enregistrement de parole neutre: il était demandé à chaque locuteur de répéter quatre phrases (de 10 secondes, environ) d'une manière neutre 30 fois chacune. Seules les deux premières phrases (d'environ 4 secondes) ont été utilisées dans l'expérience. Ces deux phrases sont: "Sa voisine est inimitable" et "Bien sûr, je connais son nom"" (phrases extraites du corpus de phrases phonétiquement équilibrés de Combescure).
b) Enregistrement de parole émotionnelle et fatigue. Notons qu'il est difficile d'obtenir des données authentiques sur l'expression spontanée des émotions et de la fatigue. Beaucoup d'études ont été faites avec l'aide d'acteurs, et il n'est pas évident d'obtenir spontanément des "émotions" de la part des sujets, dans de bonnes conditions d'enregistrement. Pour mettre les locuteurs en condition, les phrases cibles mentionnées ci-dessus ont été incluses dans des textes, suggérant l'émotion désirée (joie, colère) et la fatigue.

La base de données de parole Imitation a été constituée en enregistrant 24 locuteurs (11 hommes et 13 femmes) âgés de 25 à 50 ans, qui ont tenté d'imiter deux locuteurs de référence (un homme pour les hommes et une femme pour les femmes). Les locuteurs ont été enregistrés en trois sessions. L'intervalle entre le premier et le second enregistrement était de 15 jours, et celui entre le second et le troisième était d'un mois. Les locuteurs ont été enregistrés durant leurs heures de travail, à travers le réseau téléphonique interne d'une université.

Pour chaque enregistrement (après une phase d'entraînement), le locuteur entendait l'énoncé d'un des deux locuteurs de référence (d'une durée de 4 secondes), au téléphone avec un interface PC- téléphone. Ensuite, il lui était demandé de répéter le même énoncé en imitant le locuteur de référence. 10 énoncés de chaque locuteur ont été enregistrés.

3. VÉRIFICATION DU LOCUTEUR PAR LES AUDITEURS HUMAINS

Trois tests ont été menés. Le premier test (test I) est un test de discrimination antre parole neutre et parole chargée d'émotion avec identification des émotions. Le second test (test II) comprend une phase de vérification du locuteur sur les voix neutres, et une phrase de vérification du locuteur sur les voix émotionnelles. Le test III est un test de vérification du locuteur sur les voix imitées. Les tests I et II ont été effectués dans une salle de cours. Les auditeurs étaient des étudiants de second et troisième cycle en Phonétique. Ils ont écouté les paires de stimuli diffusés à l'aide d'un magnétophone d'une qualité convenable.

3.1. TEST D'IDENTIFICATION DE L'ÉMOTION (TEST I)

20 étudiants en linguistique ont écouté deux répétitions de chaque stimulus de la Base I et ils ont eu à juger les stimulus comme neutres ou non (choix binaire), et si le stimuli était jugé comme non neutre, l'auditeur devait indiquer les intentions du locuteur (choix forcé: joie, colère ou fatigue).
Résultats: Ce test montre que dans 70% des cas les auditeurs étaient capables d'identifier correctement l'émotion simulée du locuteur (joie, colère ou fatigue). De tels résultats indiquent que les locuteurs ont réussi environ 2 fois sur 3 à transmettre l'émotion désirée.
Nous en avons conclu que notre base de données permettait la conduite des expériences décrites ci-dessous.

3.2. VÉRIFICATION DU LOCUTEUR EN PAROLE ÉMOTIONNELLE (TEST II)

84 auditeurs, ne connaissant pas les locuteurs, ont participé à ce test. Il était expliqué aux auditeurs qu'ils allaient entendre 52 paires de stimuli et qu'ils devaient juger si les deux énoncés de chaque paire étaient produits ou non par le même locuteur. Ce test a été conduit en deux parties:
a) Les deux stimuli de chaque paire entendus par les auditeurs étaient "neutres".
b) L'un des deux stimulus était "neutre" et l'autre 'émotionnel".

Après une analyse des réponse données par ces auditeurs, seules les réponses de 69 d'entre eux ont été soumises à une analyse statistique. Les 15 locuteurs restants ont été considérés comme des auditeurs non coopératifs, soit parce qu'ils n'ont pas pris de décision sur la feuille de réponse, ou soit parce qu'il semblait répondre aléatoirement (une taux d'erreur de plus de 50%, etc.).
Résultats: Les résultats des tests II-a et II-b sont résumés dans le tableau 1 (voir commentaires plus bas).

3.3. VÉRIFICATION DU LOCUTEUR EN PAROLE IMITÉE TÉLÉPHONIQUE (TEST III)

Pour ce test, 20 auditeurs ont entendu, à travers un casque, sur ordinateur, 46 paires de stimuli lus extraits de la base de données d'imitation. Le test a été réalisé en laboratoire sur un PC 486 et une carte DSP EVM C30. Les sujets donnaient leur réponse en utilisant le clavier.

3.4. RÉSULTATS

Les résultats des tests II et III sont résumés dans le tableau 1:

	FR	FA	(FR+FA)/2
Test II-a	11%	13%	12%
Test II-b	48%	20%	34%
Test III	6%	26%	16%

FR = faux rejets FA = fausses acceptations

Tableau 1. Résultats de la vérification du locuteur par écoute de parole neutre (II-a), émotionnelle (II-b) et imitée (III).

Les tests II a et II b montrent une augmentation de 37% (de 11% à 48%) des faux rejets (FR). Les variations acoustiques de la parole dues à l'état émotionnel du locuteur réduisent donc l'aptitude des auditeurs à distinguer entre les énoncés du même locuteur produits dans deux états émotionnels et physiques différents. Ce résultat peut être comparé à celui obtenu par notre SVAL (section suivante).

4. VÉRIFICATION DU LOCUTEUR PAR LA MACHINE

La première version de notre SVAL, appelée SPREC0 [3] est élaborée à partir de coefficients LPC cepstraux et d'un algorithme de DTW[7]. Les LPCC donnent une description très efficace du spectre à court

terme pour des algorithmes DTW de vérification du locuteur [6] [8] [9] [10]. Les énoncés ont été paramètres par une analyse LPC d'ordre 12 (utilisant une fenêtre de Hamming de 30 ms toute les 20 ms).

Deux méthodes de calcul de distance ont été évaluées. Soit une séquence de référence (de longueur I) $\{X_i\}$ $1 \le i \le I$ de vecteurs de paramètres de dimension M, $X_i = (x_{i1}, ..., x_{iM})$, et la séquence de test $\{Yj\}$ $1 \le j \le J$, $Y_j = (y_{j1}, ..., y_{jM})$, et deux distances locales d1 et d2 définies par :

$$d1(X_i, Y_j) = \sum_{m=1}^{M} (x_{im}-y_{jm})(x_{im}-y_{jm}) \qquad (1)$$

$$d2(X_i, Y_j) = [(X_i-Y_j)W^{-1}(X_i-Y_j)']^{1/2} \qquad (2)$$

où W est la matrice de covariance générale obtenue pour chaque locuteur en utilisant T énoncés de référence de ce locuteur:

$$W = \frac{1}{T}\sum_{t=1}^{T} W_t \qquad (3)$$

W est également la moyenne de la matrice de covariance W_t de chaque énoncé:

$$W_t = \frac{1}{I}\sum_{i=1}^{I} (X_i-\mu)(X_i-\mu) \qquad (4)$$

avec

$$\mu = \frac{1}{I}\sum_{i=1}^{I} X_i \qquad (5)$$

Les deux bases de données de parole émotionnelle et imitée ont été utilisées par notre SVAL pour comparer les résultats en vérification du locuteur obtenus par les auditeurs humains (voir Tableau 1) et par la méthode automatique (voir Tableau 2).

Les énoncés neutres ont été divisés en deux groupes: le groupe de références (10 énoncés) et le groupe de tests, et tous les énoncés émotionnels ont été utilisés comme énoncés de tests. En utilisant le EER et les énoncés de références, un seuil de décision et une matrice de covariance générale ont été obtenus pour chaque locuteur. Les deux distances décrites ci-dessus ont été utilisées dans toutes les expériences effectuées sur la base de données de parole émotionnelle. La distance globale cumulée au cours de la programmation dynamique a été comparée au seuil de décision pour déterminer l'acceptation ou le rejet du locuteur. La vérification du locuteur par notre SVAL sur la base de données de parole imitée n'a été effectuée que par la première méthode de mesure de distance (eq. 1).

Les seuils de décision pour chaque locuteur ont été calculés a priori en utilisant les énoncés de références et a posteriori en utilisant le critère EER. Le tableau 2 montre les résultats des deux expériences.

seuils de décision	état du locuteur	distance	(FR+FA)/2
a Posteriori *	neutre	d1	2%
a Posteriori *	neutre	d2	3%
a Priori	neutre	d2	9%
a Priori	joie	d2	44%
a Priori	colère	d2	48%
a Priori	fatigue	d2	31%
a Priori	joie	d1	39%
a Priori	colère	d1	44%
a Priori	fatigue	d1	26%
a Posteriori	joie	d2	27%
a Posteriori	colère	d2	32%
a Posteriori	fatigue	d2	25%
a Posteriori	joie	d1	27%
a Posteriori	colère	d1	33%
a Posteriori	fatigue	d1	20%
a Posteriori	imitation	d1	18%

* en utilisant seulement des réalisations de références

Tableau 2. Degré d'erreur de notre système VAL sur la parole neutre, émotionnelle et imitée.

Le tableau 2 indique que le degré d'erreur augmente beaucoup pour les états émotionnels et particulièrement la colère. La mesure de distance d1 présente un plus faible degré d'erreur que d2. Le haut degré d'erreur de vérification du locuteur dû à l'état émotionnel montre bien les limitations du système et le besoin d'études sur les variations de la parole dues à l'état émotionnel pour les tâches de vérification du locuteur.

Une comparaison entre des résultats obtenus par des auditeurs humains et par notre système sur de la parole chargée d'émotions montre un taux d'erreurs presque identique pour les deux types d'approches: 34% pour les auditeurs humains (Tableau 1) et 36% d'erreur pour le SVAL (moyenne du degré d'erreurs pour les trois émotions en utilisant la distance d1 du tableau 2). Le taux d'erreurs sur le même matériel de parole imitée était de 16% pour les auditeurs humains et 18% pour notre SVAL. Pour les voix neutres, la performance de notre système est plus élevée que celle des auditeurs humains.

5. ANALYSE SPECTROGRAPHIQUE

Dans cette partie, nous décrivons nos observations sur quelques paramètres acoustiques de la parole produite dans différents états émotionnels. L'état émotionnel est connu pour avoir d'importants effets sur la production de la parole, par une réorganisation de l'activité de tous les articulateurs, le changement de vitesse d'élocution, l'intensité, la moyenne et la plage de variation du fondamental, la tension du larynx et le mode de vibration des cordes vocales (changement des temps relatifs d'ouverture et de fermeture de la glotte, etc.

La valeur moyenne de F0, la plage de variation de F0, la durée des énoncés et l'intensité ont été

calculées sur les productions de la base des données de parole émotionnelle. La figure 1 illustre un spectrogramme et le contour de F0 des productions d'un locuteur, dans la phrase "sa voisine est inimitable", extraite de la base I (parole émotionnelle), produite de façon neutre et dans trois états émotionnels et physiques (colère, joie et fatigue).

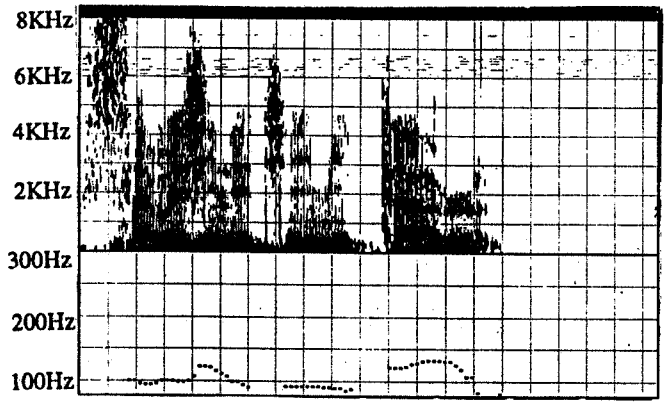


Figure 1-a) spectrogramme de la phrase "sa voisine est inimitable" dans l'état neutre.

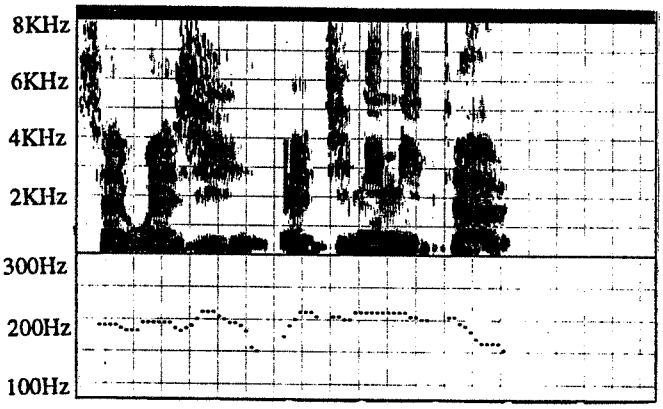


Figure 1-b) spectrogramme de la phrase "sa voisine est inimitable" simulant la joie

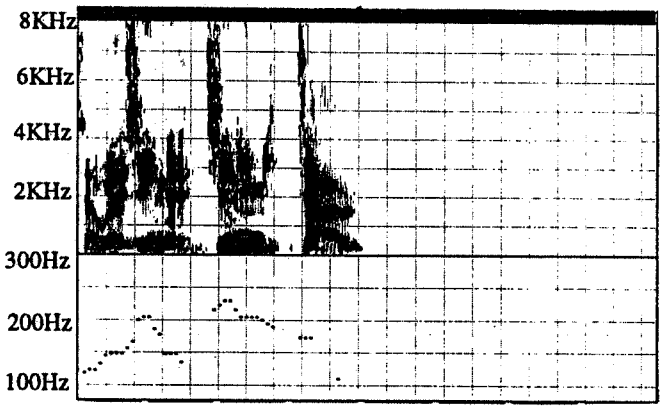


Figure 1-c) spectrogramme de la phrase "sa voisine est inimitable" simulant la colère

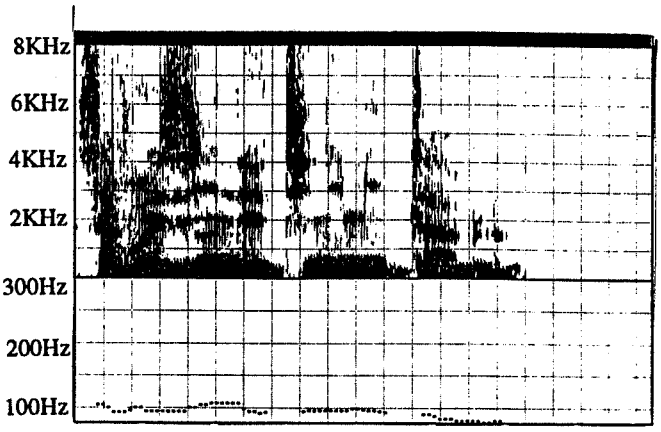


Figure 1-d) spectrogramme de la phrase "sa voisine est inimitable" simulant la fatigue

Les figures 2 et 3 illustre la différence de valeurs moyennes de F0 et de plages de variation de F0 (F0 Max - F0 Min) pour les énoncés extraits de la base I. La moyenne et la plage de F0 sont généralement les plus élevées pour la colère et les moins élevées pour la fatigue. L'intensité moyenne est généralement plus faible pour la fatigue. Ces observations sont conformes aux autres études.

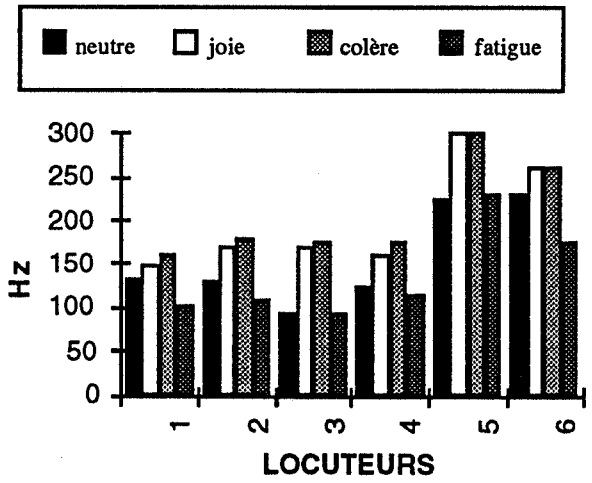


Figure 2. Valeur moyenne de la fréquence fondamentale obtenue à partir de 5 réalisations pour chaque émotion des 6 locuteurs de la base I (Locuteurs 1 à 4: hommes; locuteurs 5 et 6: femmes).

La figure 4 montre la variation dans la durée de la phrase, moyenée sur 15 énoncés pour les 6 locuteurs de la base de données de parole émotionnelle. Les durées des énoncés pour l'état neutre et pour la joie sont à peu près similaires (voir figure 4). Elles sont les plus importantes pour la fatigue.

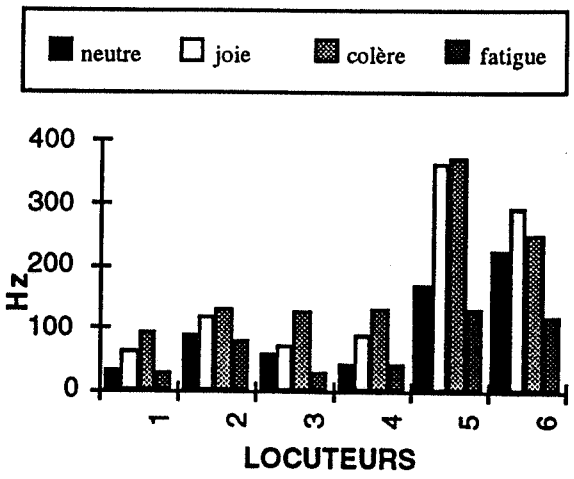


Figure 3: Plage de variation du fondamental, moyenne à partir de 5 réalisations pour chaque émotion des 6 locuteurs de la base I (Locuteurs 1 à 4: hommes; locuteurs 5 et 6: femmes).

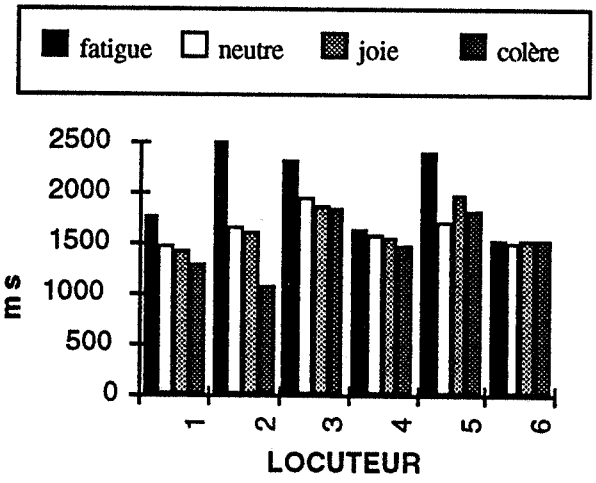


Figure 4. Durée moyenne de la phrase: "Sa voisine est inimitable" obtenue à partir de 15 réalisations pour chaque émotion produite par 6 locuteurs de la base I .

Les figures 5 et 6 montrent la fréquence moyenne du premier formant (F1) et son amplitude, en fonction de l'émotion du locuteur, pour les 6 locuteurs de la base I. Les voyelles dans la situation de joie et de colère ont des F1 plus hauts et d'amplitude plus grande. Si on considère que la hauteur du premier formant est corrélée à l'abaissement de la langue et/ou de la mâchoire, on peut considérer que le caractère fermé de la voyelle [i] est moins bien réalisé dans la parole. L'élévation de la fréquence de F1 semble être due à une tendance générale à plus ouvrir la bouche sous l'effet de l'émotion, tout au moins durant les voyelles. L'élévation observée de l'amplitude de F1 dans nos données peut avoir deux origines: (i) l'augmentation de la fréquence du premier formant elle-même (plus la fréquence de F1 est élevée, et plus l'amplitude de F1 est élevée- voir la théorie acoustique de la production de la parole de Fant), et (ii) l'augmentation de l'amplitude générale du signal, elle même due à un supplément de pression sous-glottique. On sait qu'une pression sous-glottique plus élevée, résultat de l'augmentation de la force expiratoire (liée à l'expression d'une émotion forte) a pour résultat une élévation de la pression transglottique, et en conséquent, une accélération des vibrations des cordes vocales et une élévation de l'amplitude du signal.

Une observation de l'intensité du signal des stimuli de la base de données de parole émotionnelle montre que l'intensité est grande pour la parole produite dans une situation de colère et de joie et plus faible pour l'état de fatigue. L'état de fatigue semble se traduire par une diminution générale de l'effort physiologique et un ralentissement du débit.

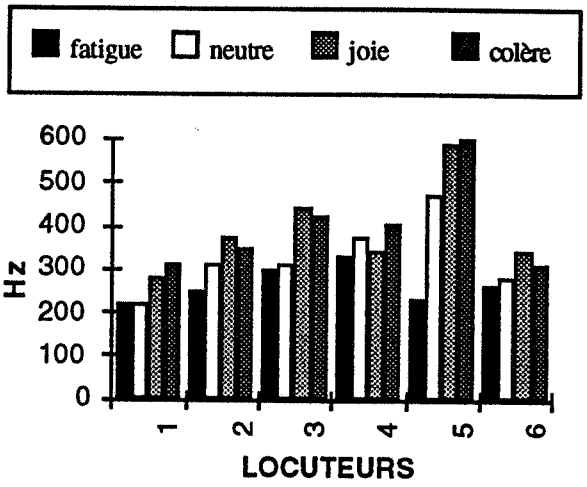


Figure 5: F1 moyen de /i/ obtenu à partir de 5 réalisations de /i/ pour chaque émotion des 6 locuteurs de la base I. (Locuteurs 1 à 4: hommes; locuteurs 5 et 6: femmes).

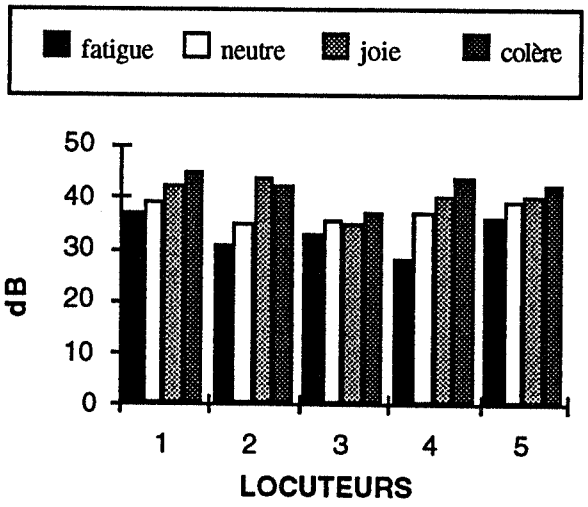


Figure 6: amplitude moyenne de F1 de /i/ obtenue à partir de 5 réalisations de /i/ pour chaque émotion des six locuteurs de la base I (Locuteurs 1 à 4: hommes; locuteurs 5 et 6: femmes).

6. CONCLUSIONS

Notre étude a confirmé l'importance de l'expression de l'émotion à la fois sur le locuteur qui modifie de façon importante sa stratégie d'articulation, et sur l'auditeur qui est alors moins capable de reconnaître l'identité du locuteur. Les tests de vérification du locuteur menés ici ont montré que les modifications dues aux émotions peuvent affaiblir la performance des systèmes de vérification du locuteur de façon relativement importante, mais comparable avec la dégradation des performances des auditeurs humains. C'est pourquoi ce type de variations doit être pris en compte dans les études sur les SVAL, et il est nécessaire de bien préciser les limites des performances de SVAL, lorsque l'élocution des locuteurs peut être modifiée par un état de fatigue anormal, ou lorsqu'il s'exprime sous le coup d'une émotion forte.

Notre étude de stimuli produits de façon neutre et dans trois états émotionnels et physiques ont montré que la moyenne et la plage de variation de la fréquence fondamentale, la durée des phrases, les variations des valeurs des formants, etc. peuvent être des facteurs pour le classement des voix produites avec différentes émotions (par exemple, F0 plus bas avec une plage de variation moindre pour la fatigue, et F0 plus haut avec une plage de variation plus importante pour la colère). D'autres mesures, telles que des modifications fines dans le mouvement de vibrations des cordes vocales sont reconnues comme importantes dans l'expression des émotions, mais elles sont difficiles à réaliser de façon automatique (il est difficile de séparer les caractéristiques de la glotte et celles du conduit vocal). Des études supplémentaires sur la vérification du locuteur et la détection automatique de l'état émotionnel sont nécessaires pour prendre en compte ces variations prosodiques et spectrales dans les SVAL et pour essayer de modifier ces variations pour conserver les bonnes performances des SVAL en dépit de ces états, si cela s'avère possible..

Cette étude à également montré que les performances des auditeurs humains et celles de notre système VAL en vérification du locuteur sont très proches lorsque les locuteurs sont dans un état émotionnel ou lorsqu'ils imitent la voix d'autres personnes. Les auditeurs humains n'arrivent pas facilement à reconnaître un locuteur, enregistré dans deux états émotionnels différents, comme étant la même personne.

REMERCIEMENTS

Nos remerciements s'adressent à l'URA 1027 du CNRS, au ministre de la Culture et des Hautes Études de l'Iran et la société PRESCOM pour leurs aides financières et matérielles. Les auteurs souhaitent remercier également tous les étudiants de l'Institut de Phonétique de Paris, et particulièrement Bertrand Lauret pour les enregistrements laborieux et les tests de perception.

RÉFÉRENCES

[1] C. E. Williams, K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates", JASA, Vol. 52, No. 4, Part 2, pp. 1238-1250, 1972.

[2] K. R. Scherer, "Emotion expression in speech and music", Music, Language, Speech and Brain, WENNER-GREN International Symposium series, Vol. 59, 1990.

[3] M.M. Homayounpour, "Vérification vocale d'identité", rapport de recherche intermédiaire. Technical Report, E.S.I.E.E., I.L.P.G.A., E.N.S.T. , 1992.

[4] M. M. Homayounpour, J.-Ph. Goldman, G. Chollet, J. Vaissière, " Performance Comparison of Machine and Human Speaker Verification", EUROSPEECH 93, pp. 2295-2298, 1993.

[5] M. M. Homayounpour, J.-Ph. Goldman, G. Chollet, "Machine vs. Human Speaker Verification", IAFP Conference, Trier, Germany, 1993.

[6] M. M. Homayounpour, G. Chollet, " A Comparison of some Relevant Parametric Representations for Speaker Verification", Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 185-188, Martigny, Suisse, Avril, 1994.

[7] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, vol. 26, NO. 1, Feb. 1978.

[8] B. Atal, "Efectiveness of Linear Prediction Characteristics of the speech wave for Automatic Speaker Identification and Verification", JASA, Vol. 55, pp. 1304-1312.

[9] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions On ASSP, Vol. 29, No. 2, April 1981.

[10] C. Bernasconi, "On Instantanous and Transitional Spectral Information for Test-Dependent Speaker Verification.

Approches Prédictives neuronales pour l'identification

T. Artières, P. Gallinari

LAFORIA-IBP UA CNRS 1095
Tour 46-00 Boite 169
Université Paris 6, 4 place Jussieu
75252 Paris cedex 05 France
gallinari@laforia.ibp.fr
artières@laforia.ibp.fr

Résumé

Nous étudions l'utilisation d'approches neuronales prédictives pour l'identification du locuteur. Après avoir mis en évidence des limitations intrinsèques à ces méthodes, nous proposons différentes directions permettant d'étendre leur champ d'application.

Mots-clés : identification du locuteur, réseau de neurones prédictif, classification de séquence, modélisation.

1. Introduction

Depuis une quinzaine d'années, de nombreux systèmes ont été testés pour l'Identification Automatique du Locuteur (I.A.L.). Les premières méthodes reposaient sur l'utilisation de mesures statiques, moyennes ou instantanées, caractéristiques du locuteur. Plus récemment, les travaux se sont tournés vers des systèmes basés sur des méthodes plus générales qui ont déjà été utilisées avec succès en Reconnaissance Automatique de la Parole [2]. Ces techniques utilisent une paramétrisation générale du signal et réalisent directement l'extraction de caractéristiques. Parmi ces méthodes on peut citer les Modèles de Markov Cachés (HMM) [16], les Modèles Autorégressif Vectoriels (VAR) [5,13] ou les Réseaux de Neurones (NN) [4,14].
L'ensemble de ces systèmes peut être divisé en deux grandes familles, selon qu'ils minimisent directement ou non une approximation du taux d'erreur en classification. Les réseaux ou les classifieurs statistiques, entraînés à produire des probabilités a posteriori des classes font partie de la première famille, alors que les HMMs, les VARs, ou les réseaux prédictifs, entraînés selon le Maximum de Vraisemblance font partie de la seconde.

La minimisation directe d'une approximation du risque moyen focalise l'apprentissage sur ce qui différencie les locuteurs, elle peut conduire à des systèmes très performants sur des grandes bases [4]. Ces systèmes présentent des inconvénients : ils nécessitent d'être réentraînés chaque fois qu'on modifie la base, les algorithmes d'apprentissage discriminant (e.g. Maximum d'Information Mutuelle) deviennent rapidement difficiles à mettre en oeuvre lorsque le nombre de classes augmente. Cela pose problème en IAL où le nombre de locuteurs est variable et n'est pas limité.
Les approches appartenant à la seconde famille visent à construire un modèle de la distribution de probabilités de chaque classe indépendamment des autres et sont donc incrémentales. Leur principal défaut tient au critère d'apprentissage qui reflète mal le but poursuivi et conduit à modéliser des informations totalement inutiles pour la discrimination entre locuteurs.

Nous étudions ici la mise en oeuvre de réseaux prédictifs, qui appartiennent donc à cette deuxième famille, pour l'IAL. Nous présentons ces systèmes dans le §2 et mettons en évidence des limitations majeures dans le §4. Nous proposons dans les §5 et §6 deux directions permettant d'améliorer leur utilisation et leurs performances pour le problème étudié. Toutes les expériences ont été réalisées sur la base de données TIMIT, présentée dans le §3.

2. Approche Prédictive.

Approcher le problème de l'identification par des méthodes basées sur la modélisation serait optimal avec une modélisation parfaite. Dans la réalité, les modèles utilisés sont structurellement loin d'être parfaits et constituent des approximations grossières des systèmes modélisés, les algorithmes d'apprentissage (Estimation Maximisation pour les HMMs, Descente de Gradient pour les NNs) ne permettent généralement que de trouver des solutions localement optimales.
Grenier a été l'un des premiers à utiliser la prédiction pour faire de l'identification en utilisant des modèles linéaires VAR [7]. Il considère alors la suite des trames, en l'occurrence des vecteurs LPC, comme la réalisation d'un processus stochastique vectoriel qui sera donc représentable par ses différents moments. Sa méthode est justifiée particulièrement dans le cas de processus stationnaire du second ordre. Ces modèles ont toutefois donné de bonnes performances sur de grandes bases de données peu bruitées [5,13].
Les réseaux de neurones prédictifs constituent une extension des techniques VAR. Bien plus que ces derniers, ils permettent de mettre en oeuvre des modèles dont la complexité et la structure doivent en théorie pouvoir s'adapter à la tâche. Nous étudions par la suite des problèmes de base concernant leur mise en oeuvre en IAL. Il s'agit des liens entre la complexité et l'adéquation du modèle, de la nature de l'information modélisée et de la possibilité d'introduction de connaissances a priori.

2.1 Réseaux prédictifs

Les réseaux prédictifs utilisés ici sont des perceptrons multicouches (MLP), à une couche cachée, entraînés à produire la trame à l'instant t en fonction d'un contexte. Pour celui-ci, nous avons pris après plusieurs essais les trames aux instants t-2 et t-1, qui constitueront donc les entrées du réseau.
Utiliser un réseau prédictif pour modéliser le locuteur revient à supposer le signal autorégressif stationnaire, c'est à dire régi par l'équation :

$$X_t = F(C_t) + \varepsilon_t \qquad \forall\, t = 1...T \qquad (1)$$

où $X_1^T = \{X_1,...,X_T\}$ est une séquence de vecteurs de paramètres, C_t est le contexte de X_t, ε_t est un bruit gaussien i.i.d., F est indépendante de t. L'apprentissage consiste à estimer les paramètres de F. Nous confondrons implicitement dans la suite la variable aléatoire X_t et sa réalisation dans la séquence de parole.
Si l'on fait l'hypothèse que le bruit suit une loi gaussienne $N(\mu,\Sigma)$, on peut donner une interprétation probabiliste au comportement du réseau. La probabilité conditionnelle de X_t sachant $F(C_t)$ est :

$$P(X_t/F(C_t)) = P(\varepsilon_t) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2} D(X_t\text{-}F(C_t))]$$

avec:

$$D(X_t\text{-}F(C_t)) = [X_t\text{-}F(C_t)\text{-}\mu]^t\ \Sigma^{-1}\ [X_t\text{-}F(C_t)\text{-}\mu] \qquad (2)$$

2.2 Probabilité d'une séquence

Considérons un modèle autorégressif d'ordre p et une séquence de parole X_1^T et prenons

$$C_t = X_{t-p}^{t-1} = (X_{t-p},...,X_{t-1})$$

On peut calculer la probabilité de la séquence X_1^T par:

$$P(X_1^T) = [\prod_{t=p+1}^{T} P(X_t/F(C_t))] .P(X_1, X_2,..., X_p)$$

En général, on élimine la partie initiatrice du processus prédictif, $P(X_1, X_2,..., X_p)$.
soit:

$$P(X_1^T) \propto \prod_{t=p+1}^{T} P(X_t/F(C_t))$$

et

$$-\log(P(X_1^T) \propto \sum_t [D(X_t\text{-}F(C_t)) + \log(|\Sigma|)]$$

En faisant l'hypothèse classique $(\mu, \Sigma) = (0,I)$, on obtient:

$$D(X_t\text{-}F(C_t)) = \|X_t - F(C_t)\|^2$$

$$-\log(P(X_t/F(C_t)) \propto \|X_t - F(C_t)\|^2$$

et:

$$-\log(P(X_1^T)) \propto \sum_t \|X_t - F(C_t)\|^2$$

Sous cette hypothèse, maximiser la probabilité de la séquence est équivalent à minimiser la somme des erreurs quadratiques.

3. Base de test

Tous les tests que nous présentons par la suite ont été réalisés sur une petite base de données constituée de quinze locutrices du dialecte dr1 de TIMIT. Celle-ci étant très propre et enregistrée en une seule session, la reconnaissance est facile pour des segments de parole suffisamment longs. Nous donnons des résultats sur des segments courts en fonction de la durée. Dans toutes les expériences, les 5 phrases SX de TIMIT ont été utilisées pour l'apprentissage, et les 5 autres (2 SA et 3 SI) pour le test. Les données utilisées par les réseaux de neurones sont des vecteurs de paramètres issus d'une analyse LPCC à l'ordre 16, effectuée sur des fenêtres de 25,6 ms, avec un déplacement de 10 ms, après échantillonnage (16000 Hz) et préaccentuation. Le temps de parole correspondant à une séquence de n trames est donc de $\frac{n}{100}$ secondes. Les phrases ont des durées comprises entre 2 et 5 secondes.

4. Complexité des modèles et identification

Les réseaux que l'on utilise prennent deux trames en entrée pour en prédire une. Ils ont tous 32 cellules en entrée et 16 en sortie.
Quand on fait varier la complexité des modèles sur une plage raisonnable, en augmentant par exemple le nombre de cellules cachées, on peut observer le comportement suivant :

- l'erreur moyenne décroît en apprentissage et en généralisation quand on augmente la complexité des modèles.
- les performances en classification passent par un maximum pour ensuite décroître (figure.1).

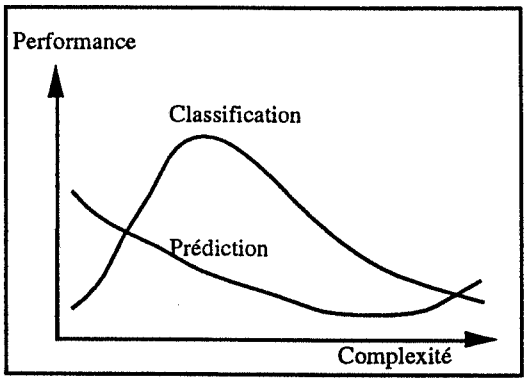


Figure 1 : comportement typique des réseaux prédictifs en fonction de leur complexité : l'optimum pour la classification apparaît bien avant le phénomène de sur-apprentissage où l'on voit l'erreur euclidienne augmenter en généralisation.

La figure 2 et la table 1 illustrent ce phénomène. La figure 2 donne les performances de trois réseaux de complexité croissante, en fonction de la longueur du test. Le meilleur

réseau en classification n'est pas celui qui offre les meilleures performances de généralisation en terme d'erreur euclidienne (table 1).
Les résultats de la table 1 montrent bien qu'il ne s'agit pas ici d'un simple phénomène de sur-apprentissage qui proviendrait de modèles trop complexes ou de données insuffisantes. C'est le critère d'apprentissage qui ne reflète pas la tâche. Au-delà d'une certaine complexité, les modèles continuent à mieux apprendre le phénomène sans rien apporter de plus pour l'identification. Des résultats similaires ont été observés par [8].

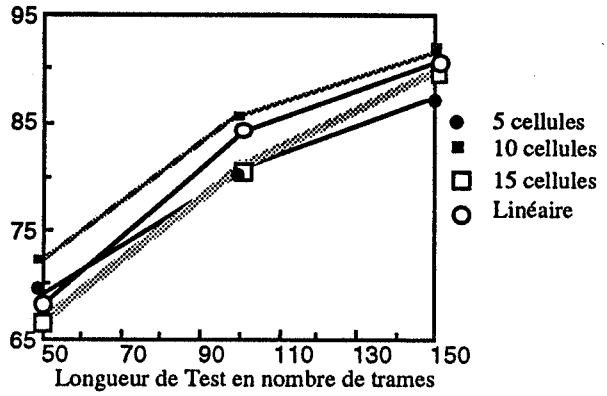


Figure 2 : performance en Identification de réseaux prédictifs en fonction du temps de parole. Les trois courbes correspondent à des couches cachées de 5, 10 et 15 cellules. "Linéaire" correspond à un réseau totalement linéaire sans couche cachée.

Complexité	5	10	15
Apprentissage	0,097	0,067	0,056
Généralisation	0,109	0,076	0,061
Séparation	0,150	0,102	0,070

Tableau 1 : erreurs euclidiennes moyennes de modèles de différentes complexités (5, 10, 15 cellules cachées) en apprentissage et en test. "Séparation" donne l'erreur moyenne d'un modèle sur les données des autres locuteurs.

Ces modèles sont donc intrinsèquement limités pour réaliser de la classification. Pour améliorer leur performance on peut suivre deux voies. La première consiste à rendre le modèle plus adéquat pour représenter le phénomène. On peut par exemple reconsidérer les hypothèses implicites dans la décomposition précédente : stationnarité, hypothèse de bruit blanc i.i.d. pour l'erreur de prédiction, ordre du modèle. Une meilleure adéquation ne se traduit pas forcément par une augmentation de la complexité. Message et identité sont extrêmement liés, on essaiera ici de mieux les modéliser simultanément. On réalisera un compromis en terme de complexité sur les modèles pour, de même que plus haut, choisir ceux donnant les meilleures performances en identification.
L'autre voie consiste à séparer dans la mesure du possible l'information locuteur du reste du signal [3].
Ces améliorations peuvent être combinées, par soucis de simplicité, nous les explorons séparément par la suite.

5. Modélisation

5.1 Ajuster la complexité des modèles

Plus le modèle sera complexe, plus il aura tendance à apprendre des informations propres à sa base d'apprentissage et inutiles pour la modélisation du signal ou la classification. On peut améliorer la généralisation en biaisant les modèles trop complexes. Une façon simple pour ce faire est de modifier la fonction de coût en introduisant un terme de régularisation. Les expériences suivantes montrent les répercussions sur les performances en classification. Nous avons utilisé comme fonction de coût :

$$C = C_1 + C_2$$

avec

$$C_1 = \sum_t \|X_t - F(C_t)\|^2$$

$$C_2 = \sum_{i,j} \|w_{ij}\|^2$$

où les w_{ij} sont les poids du réseau.

Le terme C_2 tendra à lisser la solution trouvée par le système; d'autres choix auraient pu être faits pour ce terme [9,18].

		Nbre de Trames de Test			
Nb cellules cachées	Decay	50	100	150	Phrases
5	Non	69,3	80,4	87,2	93,3
5	Oui	69,8	81,8	88,6	97,3
10	Non	72,4	85,4	90,7	93,3
10	Oui	73,7	86,4	92,2	97,3
15	Non	66,5	80,6	89,5	84,0
15	Oui	72,1	85,2	92,1	96,0

Tableau 2 : performance de différents réseaux avec utilisation optionnelle d'un terme de décroissance (decay en colonne 2) en apprentissage. L'erreur est supposée gaussienne de moyenne nulle et de matrice de covariance Identité.

Les résultats du tableau 2 montrent l'augmentation de performances. En accord avec §4, elle est d'autant plus sensible que le réseau est complexe. Ces résultats montrent toutefois que même pour des complexités faibles l'ensemble d'apprentissage introduit un biais dans la solution qui nuit aux performances du système, et qui est partiellement compensée par le terme de décroissance.

5.2 Modélisation du bruit

Un réseau prédictif d'ordre 2, entraîné à minimiser l'erreur quadratique :

$$C = \sum_t \|X_t - F(X_{t-2}, X_{t-1})\|^2$$

approxime la fonction F_{opt}:

$$F_{opt}(x,y)= \int_z z.p(X_t=z/X_{t-1}=y, X_{t-2}=x)dz$$

où l'intégrale porte sur l'espace des données. La moyenne, sur le signal appris, des erreurs de prédiction d'un réseau est une estimation de la somme suivante qui est nulle :

$$\int_{X_t} [F_{opt}(X_{t-1},X_{t-2})-X_t]p(X_t/X_{t-1},X_{t-2})dX_t$$

En pratique, sur une base de données de taille limitée, ce n'est pas le cas et il peut être utile de l'estimer.
Par ailleurs, les hypothèses sur la dispersion du bruit sous jacentes à l'emploi de l'erreur euclidienne sont clairement trop restrictives. Le tableau 3 montre l'augmentation de performances que l'on peut attendre d'une meilleure modélisation du bruit pour un réseau à 10 cellules cachées. Des résultats similaires ont été obtenus avec d'autres architectures. Un bruit an-isotrope correspondant à une matrice de covariance pleine augmente nettement les performances. L'utilisation conjointe d'un terme de régularisation augmente là aussi les résultats.

Bruit	Decay	50	100	150	Phrases
(0,I)	non	72,4	85,4	90,7	93,3
(μ,D)	non	72,2	84,9	91,8	97,3
(μ,Σ)	non	75,4	88,6	93,5	97,3
(μ,Σ)	oui	78,0	89,2	93,5	100,0

Tableau 3 : comparaison de différentes modélisations du bruit sur des réseaux à 10 cellules cachées. (0,I) est un bruit blanc, μ indique que la moyenne est estimée, D et Σ indiquent respectivement que la matrice de covariance est diagonale ou pleine.

L'hypothèse d'indépendance des erreurs est elle aussi abusive, [15] ont proposé une méthode modélisant la corrélation des erreurs de prédiction, conduisant à une amélioration de leurs résultats.

5.3 Modèles Multi-états.

Les modèles précédents considèrent le signal comme stationnaire. On peut facilement supposer que quelle que soit la nature de l'information locuteur, cela ne correspond pas à la réalité. De même que dans les Modèles de Markov, on peut se rapprocher de celle-ci en faisant une modélisation stationnaire par morceaux.
Pour cela chaque locuteur sera représenté par plusieurs réseaux modélisant chacun une partie de son élocution. Ces réseaux jouent le rôle des états d'un modèle de Markov, chacun étant dédié à une partie de l'espace des données du locuteur. Nous considérons des modèles ergodiques avec des transitions équiprobables. En effet, l'apprentissage de ces dernières n'améliore pas les performances d'un système d'IAL [12,17], du moins lorsque le nombre d'états est limité.

Un modèle se compose donc de N réseaux MLP implémentant chacun une fonction différente des entrées F_i, i = 1..N. Ces modèles sont initialisés ici à partir d'un clustering des données. Il est clair que l'initialisation est importante et que d'autres solutions peuvent être envisagées. L'optimisation est ensuite réalisée simultanément sur la segmentation du signal par les modèles et sur les paramètres de ces derniers. L'apprentissage est réalisé de la façon suivante : à chaque présentation d'un exemple, les erreurs de prédiction des N réseaux sont calculées, mais seuls les poids du réseau donnant l'erreur minimale sont modifiés par rétropropagation du gradient. Les caractéristiques du bruit sont recalculées par état, après chaque passage de toute la base d'apprentissage.
De même, en reconnaissance, le score du modèle d'un locuteur est calculé le long du chemin conduisant à l'erreur totale minimale. Puisque les modèles sont ergodiques, en notant F_i la fonction implémentée par le réseau n° i du modèle, le score d'un modèle (F_1, ..., F_N) peut s'écrire:

$$Min_{e_1...e_T}\left[\sum_t D_{e_t}(x_t-F_{e_t}(C_t))+\log|\Sigma_{e_t}|\right]$$
$$=\sum_t Min_i[D_i(x_t-F_i(C_t)) + \log|\Sigma_i|]$$

où $e_1...e_T$ est un chemin dans le modèle, avec

$\forall t = 1...T \quad e_t \in \{1,...,N\}$ et D_i est défini par (2).

Cela correspond à une approximation du maximum de vraisemblance [10]. Notons que l'optimisation du critère classique pourrait être réalisée aussi simplement en optimisant sur l'ensemble des chemins.

nb Réseaux	nb cell cachées	Bruit	nb Param	50 trames	100 trames	150 trames	Phrases
1	10	(0,I)	506	72,4	85,4	90,7	93,3
1	10	(μ,Σ)	650	75,4	88,6	93,5	97,3
1	5	(0,I)	261	69,3	80,4	87,2	93,3
3	5	(0,I)	783	74,8	86,7	91,0	100,0
3	5	(μ,D)	879	76,3	88,3	93,6	100,0
3	5	(μ,Σ)	1215	80,8	92,7	96,0	100,0

Tableau 4 : comparaison des performances de modèles à un état et de modèles ergodiques à 3 états avec différentes modélisations du bruit. Les notations concernant le modèle de bruit sont les même que précédemment. La quatrième colonne donne le nombre de paramètres de chaque modèle.

Le tableau 4 montre les performances d'un ensemble de trois réseaux à cinq cellules cachées. Il y a une nette augmentation de performances par rapport à un seul réseau du même type, mais également par rapport à des réseaux uniques plus complexes. Cela est vérifié pour différentes modélisations de l'erreur.
Le nombre de paramètres indiqué dans le tableau ne permet pas de comparer directement la complexité des modèles à un et plusieurs états. Par exemple, les modèles à 3 états et erreur N(0,I) donnent de moins bons résultats que les

réseaux à 10 cellules cachées avec une erreur $N(\mu,\Sigma)$, avec toutefois plus de paramètres.
Utiliser plusieurs états ne fait pas qu'augmenter la complexité du système mais correspond à une meilleure adéquation du modèle. Là aussi, il faut trouver un compromis en terme de complexité du système global. Les meilleures performances ont été atteintes ici avec des modèles à cinq cellules cachées. L'augmentation de performances pour l'identification ne traduit que le fait que la modélisation est améliorée.

Dans la section suivante, nous revenons aux systèmes de base à un état et modélisation de l'erreur N(0,I), pour examiner quel type d'information a priori sur la tâche il est possible d'introduire dans les modèles.

6. Incorporation de Connaissances à priori.

Nous avons vu dans les paragraphes précédents qu'il est possible, malgré les problèmes évoqués au §4, d'augmenter la performance d'un système basé sur la modélisation en utilisant des modèles plus adéquats tant au niveau de la modélisation de l'erreur que de l'hypothèse de stationnarité, ou encore de la corrélation des erreurs.
Une autre possibilité est d'y d'incorporer de la connaissance a priori. Cette connaissance peut être très variée et intégrée à tous les niveaux, de l'apprentissage à la reconnaissance. Nous présentons dans la suite deux approches, l'une basée sur le filtrage du signal, l'autre sur l'extraction de caractéristiques. Toutes deux ont pour but de détecter les trames "discriminantes", c'est à dire riches en information sur le locuteur, et présentent l'avantage d'introduire de la discrimination tout en conservant au système global son caractère incrémental. La première intègre cette information au niveau de l'apprentissage, la seconde en reconnaissance.

6.1 Modèle de locuteur moyen

Différentes approches ont été étudiées dans le passé pour séparer les informations relatives au locuteur et à la parole. Tishby [17], propose un modèle formel de factorisation de ces deux composantes du signal de parole. Nous proposons ici d'utiliser un modèle de "Locuteur Moyen" (LM), appris sur les données de plusieurs locuteurs, avec l'idée suivante : les segments de parole bien modélisés par ce modèle n'auront pas d'intérêt pour l'identification du locuteur. L'apprentissage des modèles de locuteurs gagnerait donc à être focalisé sur les segments mal modélisés par ce modèle de locuteur moyen. Une implémentation possible de cette idée est la suivante:

- entraîner un modèle prédictif de Locuteur Moyen sur les données de plusieurs locuteurs représentatifs.
- entraîner un modèle prédictif par locuteur, en utilisant une fonction de coût modifiée, où les erreurs quadratiques sont pondérées vectoriellement par les erreurs du modèle du locuteur moyen. Une trame bien prédite par le locuteur moyen, donc peu discriminante, verra son importance diminuée dans l'estimation de la fonction de prédiction. Le tableau 5 donne les performances comparées de modèles appris avec une fonction de coût quadratique et avec une fonction de coût modifiée. Cette procédure permet un gain de performance de 1 à 2 %.

# segments	Longueur de test en trames	Réseaux à 5 cellules cachées		Réseaux à 10 cellules cachées	
		standard	LM	standard	LM
950	50	69,3	71,1	72,4	73,2
800	100	80,4	82,7	85,4	86,1
650	150	87,2	88,9	90,7	91,8
75	Phrases	93,3	93,3	93,3	93,3

Tableau 5 : Performance de l'algorithme d'apprentissage avec Locuteur Moyen (LM) comparé à l'algorithme standard.

6.2 Extraction de Caractéristiques

De nombreuses expériences ont montré que certains segments de parole étaient intrinsèquement plus discriminants que d'autres. Quelques systèmes ont été construits afin d'exploiter cette connaissance pour améliorer des méthodes classiques, soit en sélectionnant des trames dans une séquence de parole [1,11], soit en leur donnant une importance relative dans la prise de décision finale [6,16]. Ces systèmes ont été testés sur des tâches peu difficiles ou en introduisant la connaissance a priori "à la main".
L'approche que nous proposons vise à prendre en compte la qualité discriminante des trames de façon automatique sur une tâche complexe. Le but est d'améliorer les performances d'un système existant en analysant son comportement. Nous partons de l'hypothèse qu'il est possible de caractériser les erreurs faites par ce système et, dans une certaine mesure, de les prévoir. Le système initial utilisé ici est un système de réseaux prédictifs, mais pourrait être basé sur des Modèles de Markov ou tout autre type de modèles.
Une trame ou un segment de parole est utile pour le système si cette trame ou ce segment contribue de façon significative à l'identification. Une fonction capable de prévoir l'utilité d'une trame ou d'un segment permettrait l'extraction de caractéristiques. Plus précisément, nous recherchons à construire une fonction G(t) à valeurs dans [0,1], représentant la confiance que nous avons dans la décision locale, au niveau trame, du système. Nous avons utilisé deux implémentations pour approximer cette fonction.

- La première consiste à réaliser une partition de l'ensemble des données d'apprentissage, et à calculer pour chaque région un taux égal à la proportion de trames bien reconnues. Pour une trame donnée, la fonction G prendra pour valeur le taux correspondant à sa région.
- Pour la seconde implémentation, nous avons utilisé un réseau de neurones utilisant en entrée le contexte (X_{t-1},X_{t-2}). et entraîné, sur les données de locuteurs représentatifs, à calculer 1 si la trame X_t est bien reconnue par le système et 0 sinon.

Nous présentons ici deux utilisations possibles de cette fonction.

6.2.1 Pondération des erreurs.

Une première utilisation de cette fonction est de modifier la procédure de reconnaissance. Pour cela, on remplace la

somme quadratique des erreurs de prédictions $\sum_t \|\varepsilon_t^i\|^2$ par $\sum_t G(t) * \|\varepsilon_t^i\|^2$.

Cela permet de donner une importance relative aux trames de la séquence, correspondant à l'estimation que nous avons de son utilité. Les deux implémentations donnent des fonctions similaires et améliorent légèrement les performances (Tableau 6).

Reconnaissance	50 trames	100 trames	150 trames	Phrases
sans G	69,3	80,4	87,2	93,3
avec G	69,9	81,6	87,8	94,7

Tableau 6 : Performances de réseaux prédictifs à 5 cellules cachées avec pondérations des erreurs locales par la fonction G, implémentée par un réseau de neurones.

6.2.2 Rejet de séquences.

G peut également être utilisée pour déterminer si une séquence contient suffisamment d'information discriminante pour que la décision du système soit fiable. Nous avons choisi de prendre comme mesure de cette information discriminante la moyenne de G sur la séquence. Si cette mesure est inférieure à un seuil, la séquence est rejetée. En faisant varier ce seuil, nous avons obtenu de nombreux résultats, résumés par la figure 3. La performance du système peut ainsi être augmentée de 2 à 3 % en rejetant un nombre limité de séquences. Dans une application réelle, cette forme de rejet pourrait être utilisée pour moduler le temps de parole demandé afin d'atteindre un niveau suffisant de fiabilité.

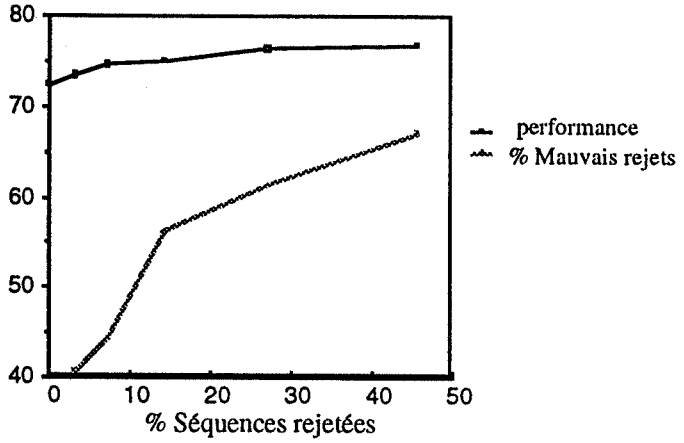


Figure 3 : rejets sur des séquences de 50 trames.
Les courbes donnent les performances de réseaux prédictifs à 5 cellules cachées et le pourcentage de séquences bien reconnues rejetées, en fonction du pourcentage total de séquences rejetées.

7. Conclusion

Nous avons présenté une approche neuronale pour l'identification du locuteur. Elle se situe dans la lignée des travaux sur les modèles autorégressifs vectoriels dont les NNs prédictifs constituent une généralisation. Les modèles neuronaux autorisent une grande variété de mises en oeuvre. Nous avons pu ainsi analyser les performances de l'approche prédictive en fonction de la complexité du modèle, et proposer plusieurs améliorations du système de base.

REFERENCES

[1] Anglade Y., Fohr D., Junqua J.C. 93 : speech discrimination in adverse conditions using acoustic knowledge and selectively trained neural networks, ICASSP 93, II 279-282.

[2] Artières T., Bennani Y., Gallinari P., Montacié C. 91 : connectionist and conventional models for free-text talker identification, Neuro-Nîmes 91.

[3] Artières T., Gallinari P. 93 : neural models for extracting speaker characteristics in speech modelization systems, Eurospeech, III - 2263-2266.

[4] Bennani Y., Gallinari P. 91 : on the use of TDNN extracted features information for talker identification, ICASSP'91.

[5] Bimbot F., Mathan L., De Lima A., Chollet G. 92 : standard and target driven AR-vector models for speech analysis and speaker recognition, ICASSP 92, II 5-8.

[6] Chang P.C., Chen S.H., Juang B.H. 91 : discriminative analysis of distorsion sequences in speech recognition, ICASSP 91, 549-552.

[7] Grenier Y. 80 : utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur, JEP 80, 163-171.

[8] Hattori H. 92 : text-independent speaker recognition using neural networks, ICASSP 92, II 153-156.

[9] Hinton G.E, van Camp D. 93 : keeping neural network simple, IEEE ICANN 93, 11-18.

[10] Levin E. 93 : hidden control neural architecture modeling of non-linear time-varying systems and its applications, IEEE Trans on NN, vol 4, Jan 93, 109-116.

[11] Li K. P., Porter J.E. 88 : normalization and selection of speech segments for speaker recognition scoring, ICASSP 88, 595-598.

[12] Matsui T., Furui S. 92 : comparison of text-independent speaker recognition methods using VQ-distorsion and discrete/continuous HMMs, ICASSP 92, 157-160.

[13] Montacié C., Le Floch J.L. 92 : AR-Vector models for free-text speaker identification, ICSLP 92.

[14] Oglesby J., Mason J.S. 91 : radial basis function networks for speaker recognition, ICASSP 91, 393-396.

[15] Petek B., Ferligoj A. 93 : Exploiting prediction error in a predictive-based connectionist speech recognition system, ICASSP 93, Vol II, 267-270.

[16] Rosenberg A.E., Lee C, Soong F. K. 90 : sub-word unit talker verification using Hidden Markov Models, ICASSP 90, 269-272.

[17] Tishby N. 88 : information theoretic factorization of speaker and language in Hidden Markov Models with application to speaker recognition, ICASSP 88.

[18] Weigend A.S., Rumelhart D.E., Huberman B.A. 91 : generalization by weight elimination with application to forecasting, NIPS, R.P. Lipmann, J.E. Moody, D.S. Touretzky eds., Morgan Kaufmann, Vol 3, 875-882.

Le développement et l'enregistrement de ICY, un corpus pour l'étude de la variabilité Intralocuteur et la Caractérisation de stYles parlés

Vincent Péan, Maxine Eskénazi

LIMSI-CNRS, BP 133, 91403 ORSAY CEDEX, FRANCE

Résumé

Cet article décrit la constitution et la mise en forme d'un corpus pour l'étude de la variabilité intra-locuteur et de styles contrastifs parlés.
Le corpus recueilli permet de comparer différents styles parlés sur les mêmes contenus linguistiques pour un locuteur donné.
Le locuteur décrit oralement des objets qui diffèrent dans deux dessins et ses descriptions sont les lieux énonciatifs où une variation phonologique peut se produire.
Les styles sont "normal", "soigné" et "lu-soigné". L'enregistrement par locuteur occupe environ 60 Moctets (audio) , ce qui représente environ 30 minutes de parole, avec une moyenne de 70 "contextes de variations phonologiques potentielles" par locuteur et par style, recouvrant: le voisement, le dévoisement, l'élision du schwa, la palatalisation, la nasalisation, et la gémination. Vingt et un locuteurs de différentes origines ont déjà été enregistrés. Les données sont en cours d'étiquetage à l'aide de techniques semi-automatiques, et devront être validées perceptivement pour confirmer le changement de style.
Cette méthodologie a aussi été utilisée pour constituer un corpus correspondant en anglais britannique.

Mots-clés: styles de parole, variabilité, caractérisation du locuteur, base de données.

Introduction

Cet article décrit le développement d'un corpus dans le but d'étudier le comportement linguistique individuel de locuteurs pour différents styles parlés. Cette étude est envisagée dans la perspective du développement de règles de transformation graphème-phonème. Nous désirons donc étudier la variation de style au niveau phonologique. Ainsi l'observation des faits se fera aux niveaux perceptif et acoustique, et la variabilité ne sera prise en compte que si elle correspond à une modification d'une unité phonologique du système par rapport à celle produite par une référence (GRAPHON) dans le même contexte. Nous parlerons alors de variabilité phonologique. La base de données comprend des locuteurs de diverses origines (et âges, sexes, etc.) et doit permettre : a) la comparaison de differents styles parlés pour un locuteur donné (ou un groupe de locuteurs) ; b) la production et la perception claires de styles de parole ; c) la production oralisée non-lue du même contenu linguistique par les locuteurs, qui consiste en des contextes prédéfinis où des variantes phonologiques peuvent se produire. Trois styles parlés sont générés ici : deux "spontanés" non-lus et un lu.
Quand on recueille des exemples de divers styles parlés, on se trouve confronté à quelques problèmes classiques. Ainsi, même s'il est préférable de collecter de la parole totalement libre, ce type de parole est très difficile à utiliser pour des études comparatives. En effet le contexte pragmatique varie constamment, et par conséquent, les styles parlés aussi (par exemple : au sein d'un même énoncé on peut passer d'un style très "normal" avec un ami, à un style plus formel pour transmettre une information importante). De plus, si un matériau linguistique spécifique doit être étudié, il peut ne pas se trouver en quantité suffisante dans une conversation non dirigée ; ce matériau ne sera probablement pas identiquement présent pour tous les locuteurs. Un corpus permettant de comparer des éléments linguistiques donnés sur divers styles de parole doit par conséquent être fondé sur de la parole provoquée. Il est alors important de conserver la parole aussi naturelle que possible et d'éliciter des styles parlés qui sont comparables. Il est clair que lors de l'étude de données de parole, l'influence du contexte de situation est significatif : il est le résultat d'un certain nombre de facteurs tels que la personnalité du locuteur, celle de l'auditeur, l'environnement immédiat [1]. Ainsi la comparaison de la parole recueillie lors d'un dialogue entre deux amis et lors d'un exposé d'une conférence a peu de sens pour notre propos : en effet comment attribuer les différences obtenues au style? Pour que les styles de parole soient comparables, le contenu et le contexte de la tâche permettant de les obtenir doivent être comparables.
La description qui suit concerne la base de données du français, le travail envisagé sur la version de l'anglais britannique a été décrite dans un précédent article [2] nous n'y reviendrons pas. Un autre enregistrement a commencé à l'Université de Provence dans l'environnement d'enregistrement ACCOR (mesures palatalographiques...).

Développement de la base de données: Contextes prédéfinis

Si le comportement phonologique d'un locuteur donné doit être décrit, il faut faire un choix judicieux des contextes phonologiques où la variabilité peut se produire. Les sources de variabilité phonologique en français sont nombreuses [3],[4],[5],[6] ; un choix restrictif est nécessaire. Nous avons sélectionné les variantes les plus fréquentes, présumant qu'elles seront aussi celles que nos locuteurs seront les plus à même de prononcer. Elles sont : l'élision de schwa, le dévoisement, le voisement, la palatalisation, la nasalisation et la gémination. En francais, l'assimilation est plus fréquemment a) régressive (le second phonème influence le précédent, et non le contraire comme en anglais) ; et b) entre des mots plutôt qu'à l'intérieur des mots [2],[7]. Les contextes ont été constitués par conséquent de groupes de mots porteurs de contextes de variation phonologique potentielle à leurs frontières (par exemple, "chaussure rose").

Développement de la base de données: la tâche

Des listes de variantes potentielles prédéfinies ayant été générées, il était nécessaire alors de trouver une tâche amenant le locuteur à les prononcer sans les lire et cela plusieurs fois, dans différents styles chaque fois et aussi naturellement que possible. Il a été décidé des expériences passées [8] que la tâche ne comporterait pas de scénario à apprendre, ceci pour éviter toute ambiguïté ultérieure dans la comparaison entre les styles "spontanés" et le style lu. Les styles obtenus doivent avoir un contexte commun et être clairement percus comme différents les uns des autres.

Dans le but de préserver une qualité de "spontanéité" , il a été décidé que le locuteur décrive des objets dans un "jeu des sept erreurs". La tâche consiste à décrire les différences entre deux dessins qui, à première vue, paraissent identiques, mais qui en fait, contiennent des objets différents. Le locuteur doit trouver toutes ces différences et les décrire. Pour notre but, les objets qui diffèrent dans les dessins sont descriptibles oralement par des groupes de mots prédéfinis portant les contextes phonologiques où une variation pourrait se produire (par exemple, "chaussure rose", gémination du /R/ - sur le dessin de gauche et "chaussure rouge", gémination du /R/, aussi - sur le dessin de droite). Le dessin de gauche est considéré comme le "dessin de référence" à partir duquel la description est débutée chaque fois. Le locuteur est aussi amené à donner la localisation de l'objet au sein du dessin et à le décrire complètement, avec sa couleur, par exemple. L'attention du locuteur est par conséquent très absorbée par la tâche.

Le choix de contextes phonologiques précis a été compliquée par deux facteurs : a) une contrainte syntaxique : les mots d'un groupe porteur des contextes de variabilité phonologique potentielle doivent être prononcés dans un ordre donné pour obtenir la prononciation effective du contexte ("chaussure rose", pas "chaussure" ou "chaussure qui est rose") ; b) l'image que décrit le locuteur doit entraîner une description non ambiguë (rouge qui ne pourrait pas être confondu avec orange, ou une chaussure qui n'a pas l'air d'un chausson). Les contextes phonologiques ont été choisis de façon à ne pas tous se présenter dans la même structure syntaxique [9] (tel qu'un nom suivi par un adjectif postposé, comme ci-dessus), et tels qu'ils se trouvent au sein d'un groupe syntaxique ou en frontières de deux groupes syntaxiques. Les structures sélectionnées furent: nom + adjectif ; adjectif + nom ("longue cravate") ; nom + adjectif + préposition ("fleur rouge sur") ; adjectif + nom dans une date ("le douze septembre") ; nom + conjonction + nom dans une heure ("midi et demie"). Le locuteur a été amené à donner une description aussi complète que possible des différences. De nombreux "contextes libres", contextes de variabilité phonologique potentielle qui n'ont pas été prédéfinis, ont aussi été produits, augmentant ainsi le nombre total de contextes prononcés. Les contextes d'élision de schwa, nombreux en français, ont tous été pris en contexte libre.

Il y a quatre paires de dessins.

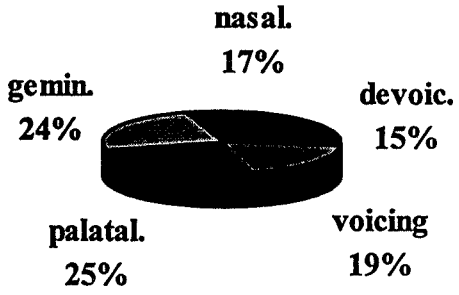


Figure 1. Contextes de variabilité phonologique potentielle prédéfinis (tout style).
Détails : 19 contextes de voisement potentiels, 9 de dévoisement, 18 de palatalisation, 2 de palatalisation + voisement, 7 de palatalisation + dévoisement, 19 de nasalisation, 27 de gémination.

La figure 1 montre qu'une quantité à peu près égale des différents types de contextes a éte prédéfinie.

Styles de parole

Une fois la tâche et les contextes définis, des essais empiriques furent tentés pour perfectionner les instructions qui pourraient amener un locuteur à produire les styles de parole sélectionnés. Les trois styles sont recueillis dans le même contexte pour pouvoir ensuite être comparés - même le "style lu" n'est

pas un style de parole en soi du fait de la nature très différente de la parole produite quand, par exemple, on lit à un enfant, ou on lit un journal à voix haute; quand on a l'habitude de lire à voix haute ou quand on ne l'a pas. La situation choisie est donc un contrat pour des enregistrements vidéo dans le but d'aider des enfants malentendants à apprendre la lecture labiale [8]. Le locuteur réalise cette tâche en trois étapes. D'abord il y a la "répétition" durant laquelle il croit que l'équipement vidéo ne fonctionne pas. Ceci fournit le style "normal" de la parole. Puis le "vrai enregistrement" est effectué (cette fois le locuteur voit que la caméra fonctionne). Ceci donne le style "soigné" de parole. La parole est ensuite transcrite et, plus tard dans la même journée, le locuteur lit la transcription en vue d'un autre enregistrement que les enfants pourront utiliser pour apprendre la lecture labiale. Ainsi la parole lue est recueillie à partir de la même tâche et du même but que les paroles non lues.

Enregistrement

Chaque enregistrement a été effectué dans une chambre sourde. Le locuteur portait un microphone "sur tête" SHURE SM10 connecté à une carte VECSYS TDS AD/DA. Il était aussi enregistré avec une caméra vidéo Sony HI-8. Le signal de parole était échantillonné à 16kHz et stocké à la fois sur DAT et sur WARM (disque optique). Durant l'enregistrement, chaque fichier contenait une description totale d'un dessin. Il y avait ainsi quatre fichiers par style x trois styles = douze fichiers par locuteur.

Mise en mémoire des données

En moyenne, l'espace mémoire occupé par un locuteur (pour tous les styles et dessins) est de 60 Moctets (variant de 50 à 80 pour un locuteur très expansif). Il y a 21 locuteurs actuellement et les enregistrements se poursuivent. Il y a 11 locuteurs et 10 locutrices d'âge compris entre 19 et 54 ans. Ils ne sont pas tous parisiens. Les prochains enregistrements élargiront la variété des locuteurs, en particulier leurs vécus linguistiques et culturels (précisés par un questionnaire).

Structure des données

Les données ont été rangées en une structure arborescente de répertoires compatible avec les moyens de stockage et le type des données que nous aimerions comparer le plus souvent. Ceci est montré Figure 2. La description d'un dessin entier a été découpée en fichiers plus petits correspondant chacun à la description complète d'un objet sur les dessins de gauche et de droite (par exemple, la comparaison de la couleur des chaussures à gauche et à droite). Diviser ainsi le fichier était facile car il y avait toujours une pause entre les descriptions des paires d'objets. Ceci donne une moyenne de 20 fichiers de parole par dessin. Chaque fichier-signal a cinq fichiers associés : la transcription orthographique, la transcription phonétique, la transcription phonétique idéale, la transcription syllabique-phonémique, le fichier étiqueté (par syllabes), et le fichier de marqueurs de discours.
Les enregistrements vidéo n'ont pas été synchronisés avec l'audio. Les marqueurs de discours sont posés actuellement dans le fichier d'étiquetage (à un niveau différent de celui des phonèmes) et dans un fichier séparé. Les différentes pauses sont aussi soigneusement transcrites.

Etiquetage

L'étiquetage est semi-automatique. Les transcriptions orthographiques sont d'abord passées à travers un module de conversion graphème-en-phonème, GRAPHON, qui a été légèrement modifié pour prendre en compte l'aspect spontané de la parole. La transcription phonétique "idéale" est alors sauvegardée, et une copie est automatiquement découpée en syllabes (à l'aide du module SYLLA, développé au LIMSI et encore modifié pour la parole spontanée). Le fichier ASCII résultant est étiqueté manuellement et, au vu du spectrogramme du signal et à l'écoute, on place toutes les frontières syllabiques, en utilisant le logiciel UNICE, et on modifie la transcription "idéale" en fonction de ce qui a été réellement dit. L'utilisation du signal acoustique appelle une remarque : il permet d'écouter des segments de parole sans signification et donc de vraiment noter ce qui a été effectivement prononcé (sans intervention du système cognitif). Le contenu du fichier syllabique corrigé est automatiquement inséré entre les marqueurs de frontières et le fichier d'étiquetage résultant est vérifié en utilisant encore UNICE.

Les contextes phonologiques

En se basant sur les 21 premiers locuteurs, nous pouvons observer le pourcentage de contextes phonologiques prédéfinis (c'est-à-dire le lieu où une variation phonologique peut se produire) réellement prononcés et le nombre de contextes libres également prononcés.

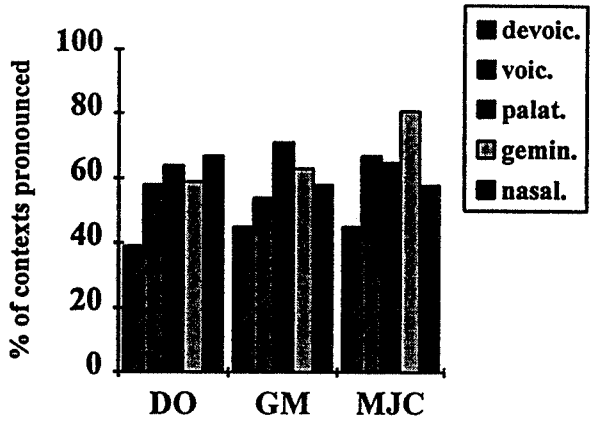


Figure 2. Pourcentage de contextes phonologiques prédéfinis actuellement prononcés (style "normal") par trois locuteurs, DO, GM, et MJC.

La Figure 3 montre le pourcentage de contextes prédéfinis prononcés par trois locuteurs choisis aléatoirement. De 40 à 80% des contextes apparaissent. Les contextes de dévoisement tendent à être moins présents que les autres contextes de variantes.

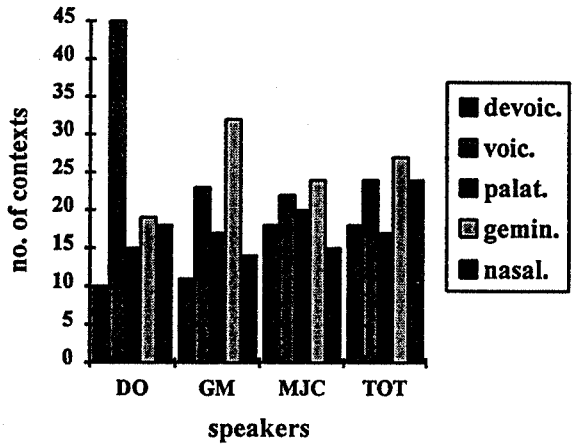


Figure 3. Contextes libres plus nombre de contextes prédéfinis par locuteur.
(style "normal")

Si nous additionnons maintenant le nombre de contextes libres au nombre de contextes prédéfinis, nous voyons que le contexte de voisement (c'est-à-dire le lieu où un fait de voisement peut se produire) est bien représenté pour tous les locuteurs. Les contextes de variantes sont assez bien représentés. Le locuteur DO, le moins expansif des trois, a prononcé un nombre plus faible de contextes prédéfinis.

Mise en forme des données

Parallèlement à l'étiquetage semi-automatique, il est nécessaire, pour les études ultérieures des données, d'effectuer un traitement sur les chaînes de caractères représentant les divers énoncés.
Ainsi nous plaçons des marqueurs de discours, des pauses, des frontières de mots et, à l'intérieur de chaque mot des frontières de syllabes (en fonction du découpage effectué par le module de règles SYLLA, donc de manière automatique et régulière).
La représentation des marqueurs de discours et des pauses par des caractères spéciaux permettra l'étude de leur distribution et de leur corrélation aux phénomènes de variabilité phonologique. Leur analyse est actuellement réalisée en parallèle par Anne Lacheret et Christel Leleu du LIMSI.
La représentation des frontières de mots est importante : elle constitue un apport d'information pour l'étude distributionnelle des phénomènes de variabilité phonologique. De plus l'information sur le caractère poly- ou monosyllabique des mots (donnée par les frontières de syllabes) apparaît comme essentielle pour le caractère distributionnel des variantes [11].
La méthodologie d'étude distributionnelle des variantes phonologiques est la suivante : nous utilisons la chaîne "idéale" générée par GRAPHON comme référence (non comme une norme) pour relier l'étude des variantes dans deux styles différents. En effet les énoncés obtenus pour un locuteur donné et deux styles distincts ("normal" et "soigné") différents, et il n'est pas possible de les comparer directement pour faire ressortir les différences de comportement phonologique. L'utilisation des chaînes "idéales" données par GRAPHON dans chaque style permet de caractériser le comportement phonologique du locuteur par rapport à un système de règles fixes dans les deux cas.
Pour un locuteur donné, les résultats de la confrontation simultanée de chaque chaîne "réelle" à sa chaîne "idéale" correspondante seront alors comparables. Cela revient en fait à déplacer le locuteur en fonction du style dans un espace dont la référence est un locuteur ayant le même comportement phonologique que GRAPHON.
Il est important de remarquer que la chaîne "idéale", dans un style donné, correspond à la sortie du module GRAPHON de la transcription orthographique de l'énoncé légèrement modifiée. En effet, pour faire apparaître les phénomènes d'élision et de réduction ('il y a' prononcé 'ya'; et 'petite', 'p'tite' ; etc.) il est nécessaire de "reconstruire" les énoncés réduits mais toujours en maintenant une régularité dans l'application de la reconstruction. Ainsi l'énoncé prononcé 'ya' par le locuteur sera transcrit 'il y a'.

Le traitement

Un module de comparaison des chaînes de caractères, développé en C, permet de faire ressortir les phénomènes d'élision, d'insertion, et de substitution entre l'énoncé phonémique réel et l'énoncé idéal. Ces phénomènes sont donnés avec une information positionnelle de leur situation dans l'énoncé, ce qui

permettra ultérieurement d'étudier leur co-occurrence libre [12] mais aussi ordonnée.
Un module permet, après la comparaison précédente, d'obtenir le type du phénomène (omission, insertion, substitution), l'élément concerné (phonème, ou autres), ses contextes immédiats gauche et droit, et son contexte lexical dans la chaîne idéale. Par exemple :

o.v/emg.w--> v <--z.imd/#xx~xw#vzx#

signifie : omission (o) de l'élément v en contexte inter-mot à gauche (emg) d'élément w et en contexte intra-mot à droite (imd) d'élément z ; le contexte lexical étant le mot polysyllabique xxxw à gauche et l'élément v étant à l'initiale du mot monosyllabique vzx.

Ainsi les études de données pourront utiliser des informations distributionnelles de plusieurs niveaux liées aux positions des phenomènes de variabilité.

Conclusion

Cet article a présenté la base de données ICY qui utilise l'élicitation pour obtenir de la parole ayant le même contenu linguistique pour trois styles de parole comparables. Une utilisation de cette base de données sera de regrouper les locuteurs ayant un comportement phonologique commun quand ils changent de styles, et de valider les résultats en utilisant la synthèse de parole. Des extensions ont déjà été planifiées. Les locuteurs feront un entretien-questionnaire durant lequel des informations importantes concernant leur "vécu" et leurs impressions de ce qu'il devait faire durant les sessions d'enregistrement seront recueillies. La coïncidence des marqueurs de discours et des variantes phonologiques est à l'étude, ainsi que la détermination d'éléments linguistiques qui sont totalement dépendants de la tâche. Dans le but de vérifier la qualité de notre base de données, nous avons aussi prévu des tests de perception pour voir si le style intentionnel du locuteur était effectivement percu par des auditeurs. Nous examinons aussi actuellement les stratégies utilisées par les locuteurs quand ils décrivent l'image.

Ce travail a été conduit en relation avec le projet CEE ESPRIT VOX Project no. 6298.
Nous tenons à remercier, Sophie Faye, Guy Chastagner, Anne Lacheret, Christel Leleu pour leur participation en diverses parties de ce travail.

Bibliography

[1] LADEFOGED P., "Three areas of experimental phonetics". London: Oxford University Press, 1967.
[2] PEAN V., WILLIAMS S., ESKENAZI M. "The design and recording of ICY, a corpus for the study of Intraspeaker variability and the Characterisation of speaking stYles". Eurospeech'93, 3rd European Conference on Speech Communication and Technology, Berlin, 21-23 Sept. 1993.
[3] ESKENAZI M. & VAISSIERE J. "Cours sur les indices acoustiques du français". LIMSI,Orsay & CNET, Lannion.
[4] MARIANI J. Introduction au traitement automatique de la parole. LIMSI. 1984.
[5] WALTER H. La phonologie du français. P.U.F. Collection Le linguiste. 1977.
[6] LIENARD J.S., Les processus de la communication parlée. Introduction à l'analyse et à la synthèse de la parole. Editions Masson. 1977.
[7] FOUCHE P. Traité de prononciation française. Ed. Klincksieck. Paris. 1968.
[8] ESKENAZI M. "Changing speech styles: strategies in read speech and casual and careful spontaneous speech". Proc. ICSLP Conference, Banff, p. 755-758, 1992.
[9] PRICE P.J., OSTENDORF M., SHATTUCK-HUFNAGEL S., and FONG C.. "The use of prosody in syntactic disambiguation". J. Acoust. Soc. Am., vol. 90, N°. 6, December 1991.
[10] PICHENY M.A., DURLACH N., BRAIDA L.D. "Speaking clearly for the hard hearing I: Intelligibility differences between clear and conversational speech". J. of Speech and Hearing Research, vol. 28, p. 96-103, March 1985.
[11] LACHERET A., "Automatic generation of phonological variation". Eurospeech' 89, Paris Sept. 1989.
[12] BERNSTEIN J., BALDWIN G., COHEN M., KAHN M., et MURVEIT H., "Co-occurrence of allophonic forms",112th meeting of the Acoustical Society of America, Anaheim, CA, 10 December 1986.

BDBRUIT
UNE BASE DE DONNEES "PAROLE" DE LOCUTEURS SOUMIS A DU BRUIT

J. ZEILIGER*; J-F. SERIGNAT*; D. AUTESSERRE**, C. MEUNIER**.

*INSTITUT DE LA COMMUNICATION PARLEE, Grenoble -
**INSTITUT DE PHONETIQUE, Aix-en-Provence - FRANCE

Résumé

La "Base de Données Parole BDBRUIT" est destinée à permettre l'étude générale des perturbations de la production de la parole, principalement au niveau de la source vocale, par le bruit environnant. Elle permettra donc l'étude de l'effet Lombard [1],[2],[3] pour le français. Un corpus a été enregistré par des locuteurs sélectionnés, soumis à des conditions de bruit variées. Cet article décrit la conception et la réalisation de cette base "parole" française.

Mots-clés: Base de données, parole, bruit ambiant, bruit blanc, source vocale, effet Lombard.

I. BDBRUIT ET SON CADRE

La conception et l'enregistrement de la base BDBRUIT sont le fruit d'une action initiée par le GDR-PRC "COMMUNICATION HOMME- MACHINE" en 1991. Un des buts de ce Groupement de Recherches qui regroupe six laboratoires, est de fournir aux chercheurs de larges bases de sons, utiles tant pour l'étude de la langue française que pour les recherches en traitement automatique de la parole. Le GDP-PRC CHM est très actif au niveau français et européen dans le domaine des bases de données et a joué un grand rôle dans la définition de standards et de procédures européens. Le GDR-PRC a notamment soutenu l'opération BDSONS (Base de Données des SONS du français) [4] et l'opération EUROM1 (Base de données européenne multilingue) [5] en confiant à l'ICP la réalisation de sa partie française. Ces deux bases respectent les recommandations issues du projet ESPRIT n°2589 SAM (Speech Assessment Methodologies and Standardization) [6] dont elles ont adopté les standards. De la même façon BDBRUIT a été réalisée dans l'esprit "SAM": l'ensemble des corpus et celui des locuteurs de BDBRUIT sont des sous-ensembles de ceux utilisés pour l'enregistrement de la partie française de la base multilingue EUROM1; la méthodologie et les outils (logiciel d'enregistrement de corpus automatisé EUROPEC [7]) exploités par l'ICP, où a été effectué l'enregistrement de BDBRUIT pour le compte du GDR-PRC, sont issus du projet "SAM".

Cette action a été menée en collaboration avec l'Institut de Phonétique d'Aix-en-Provence pour le choix des corpus et la mise au point de la méthodologie d'enregistrement.

II. LE CONTENU

II.1 LE CORPUS

Il constitue un sous-ensemble de ceux de la partie française de la base multilingue EUROM1.

Phrases:

Il s'agit de 30 phrases choisies parmi les 50 phrases "phonétiquement équilibrées" retenues pour l'enregistrement de EUROM1. Ce sont les listes n°9, 19 et 20 de dix phrases phonétiquement équilibrées mises au point par P. Combescure [8]. Afin d'éviter une lecture monotone, la ponctuation a été modifiée : on a ajouté un point d'interrogation ou un point d'exclamation à la fin de certaines des phrases de chaque liste. Ainsi il n'y a jamais plus de deux phrases successives se terminant par un point simple.

Logatomes isolés:

Les 82 logatomes sélectionnés correspondent à une organisation syllabique de type C(C)VC (combinaison Consonne ou groupe de consonnes + Voyelle + Consonne) représentant les voyelles et les consonnes du français en syllabe fermée. Ils sont répartis en cinq groupes:

variation	forme	nombre
consonne initiale	C + "il"	(20)
consonne initiale	C + "al" ou "oul"	(23)
deux consonnes initiales	CC + "il"	(5)
consonne finale	"li" + C	(19)
voyelle interne	"t" + V + "t"	(15)

Nombres:

30 nombres ont été choisis de la manière suivante :
- les chiffres de 0 à 9
- vingt nombres (entre 0 à 9999) choisis pour couvrir

les possibilités phonotactiques de l'ensemble de nombres considéré. (5ème groupe de nombres de EUROM1).

On trouvera une description détaillée de ce corpus dans une annexe du rapport final du projet SAM [9].

II.2. LES LOCUTEURS

Tous les locuteurs sont de langue maternelle française. Ils sont au nombre de 10 (5 hommes et 5 femmes) et tous originaires du bassin parisien. Ils sont représentatifs d'un français standardisé [10] pour sept d'entre eux, les trois autres ayant un léger accent parisien. Ils constituent donc un groupe relativement homogène, géographiquement et socio-culturellement : ils ont tous une formation supérieure. Ces locuteurs correspondent à ceux du groupe II de la base EUROM1.

III. CONDITIONS D'ENREGISTREMENT

III.1 NIVEAU ET TYPE DE BRUIT

Deux types de bruit ont été retenus :
- un bruit blanc
- un bruit réaliste, en l'occurrence ici : un bruit d'ambiance ("cocktail-party").

Ces deux bruits ont été extraits du CDROM "NOISE-ROM-0", produit par le "TNO Institute for Perception" à Soesterberg (The Netherlands). (Les références sur le CD original sont : signal.000 pour le bruit blanc et signal.019 pour le bruit d'ambiance. Les signaux sont reproduits sur les CDROM, avec la gracieuse autorisation de TNO, de façon à permettre toute reconstitution ou opération ultérieure).

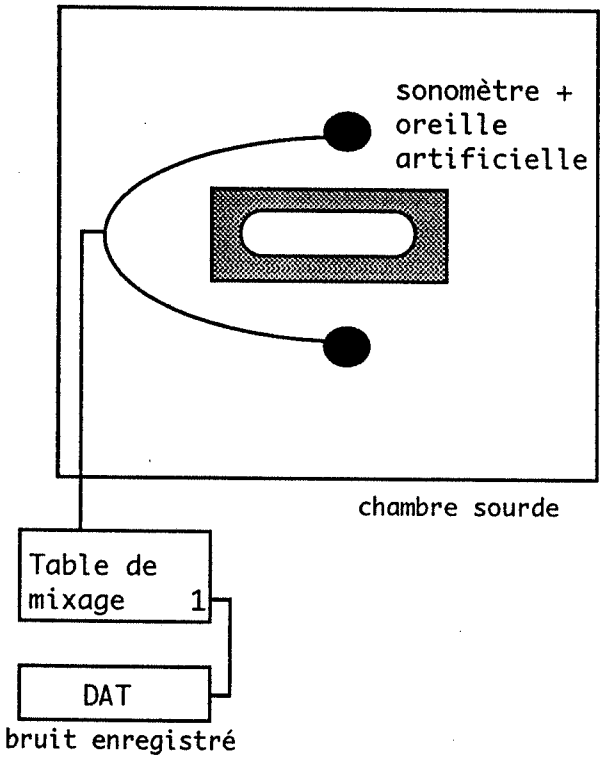


fig.1 : Principe de l'étalonnage du niveau de bruit

Les bruits ont été produits avec un niveau de 85 dB SPL au casque sur les oreilles du locuteur; le niveau ayant été étalonné à l'aide d'une oreille artificielle (B&K 4153) et d'un sonomètre (B&K 2203) (cf fig 1).

Les locuteurs ont enregistré le corpus dans cinq conditions de bruit différentes :
- milieu non bruité (standard)
- bruit blanc sans retour de sa propre voix
- bruit blanc et retour de sa propre voix
- bruit réaliste sans retour de sa propre voix.
- bruit réaliste et retour de sa propre voix.

III.2 NIVEAU DE RETOUR

Chaque locuteur n'ayant pas le même niveau de voix, le niveau de retour a été réglé sur une moyenne correspondant à un retour jugé confortable pour tous les locuteurs, fixé expérimentalement à 72 dB SPL au casque sur les oreilles du locuteur; ce niveau a également été étalonné à l'aide de l'oreille artificielle et du sonomètre, avec comme source de référence un calibrateur 1 KHz (B&K 4230) (cf fig 2).

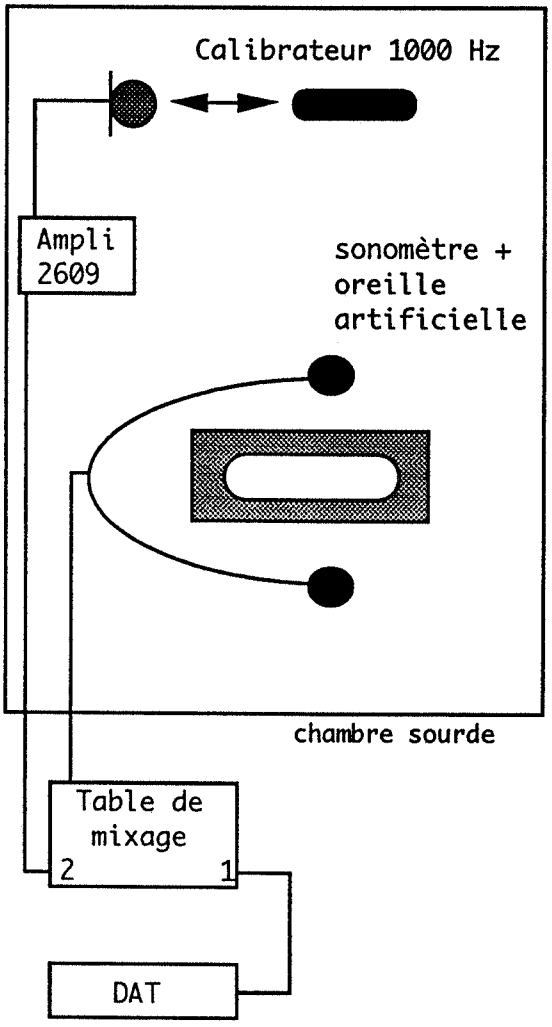


fig. 2 : Etalonnage du niveau de retour de voix dans l'écouteur.

III.3 MIXAGE

Le bruit était envoyé aux oreilles du locuteur via la première voie d'une table de mixage (Yamaha MV802). Pour les sessions avec retour de la voix du locuteur, la sortie du micro était rebouclée en utilisant la deuxième voie de la table de mixage.

III.4 MODALITES D'ENREGISTREMENT

Les enregistrements ont été réalisés en chambre sourde à l'ICP Grenoble. Pour l'ensemble de la chaîne acoustique, le niveau de bruit de fond (silence entre les mots) est à -70 dB du niveau maximum d'enregistrement pour un niveau crête du signal à environ -10 dB. Des procédures de test et de calibration, conformes aux recommandations du projet SAM [11], ont été mises en oeuvre pour assurer la qualité des enregistrements à chaque étape.

Les enregistrements ont été effectués en mode dit "continu", de façon à collecter non seulement les réalisations phoniques mais aussi les évènements extra-linguistiques dont la connaissance sera indispensable lors de l'analyse de corpus de parole naturelle. En effet, durant les pauses dites "silencieuses" entre deux phrases, des bruits sont présents : bruits respiratoires [12], bruits labiaux ou buccaux associés à la déglutition, raclements de gorge, etc...

La campagne d'enregistrement a été menée au cours de l'année 1992, avec une cinquantaine de sessions d'enregistrement (une session d'enregistrement correspondant à l'enregistrement de tous les corpus par un seul locuteur, dans une des cinq conditions prévues). Chaque session durant en moyenne deux heures, se décomposait en 1 heure d'enregistrement (locuteur dans la chambre sourde) et une heure de vérification (validité et qualité) et de sauvegardes diverses.

IV. DONNEES DISPONIBLES

IV.1 CORPUS

Les dix locuteurs ont prononcé dans chacune des cinq conditions de bruit:

3 fois tous les logatomes
5 fois tous les nombres
1 fois toutes les phrases

Tous locuteurs confondus, la base contient donc 300 phrases, 1500 nombres, 2460 logatomes, le tout disponible pour cinq conditions différentes d'enregis--trement (référence + 4 bruits).

Le tableau ci-dessous présente de manière synthétique toutes ces informations (cf fig 3).

NB: un "item" est ici un constituant de base d'un corpus, c'est donc suivant le type de données : 1 logatome ou 1 phrase ou 1 nombre...

IV.2 POPULATION DES LOCUTEURS

Les dix locuteurs retenus ont entre 20 et 50 ans et présentent les caractéristiques décrites au §II-2. Cette population est représentative d'une prononciation du français pour laquelle les opppositions de timbres vocaliques sont généralement bien différenciées, principalement en ce qui concerne les voyelles d'aperture intermédiaire (sauf "in" et "un" confondus).

Au moment de l'enregistrement trois locuteurs (2F, 1H) sur les dix ont déclaré être "fumeurs" (moyen ou gros). Comme il n'y avait pas d'incidence audible sur leur prestation, ils ont été tout de même retenus.

IV.3 VOLUME DES DONNEES

Le volume de données stockées correspondait à environ 40 à 45 Mo par session. Avec 5 conditions par locuteur et 10 locuteurs, le volume total est environ :

$$45 \times 5 \times 10 = 2200 \text{ Mo} = 2{,}2 \text{ Gigaoctets}$$

Ce volume pour la base de données BDBRUIT correspond à 4 disques CD-ROMs, dont le pressage a été effectué au cours de l'année 1993.

Les fichiers de sons y sont regroupés par type d'item, puis par type de condition d'enregistrement, pour tous les locuteurs. Le disque 1 contient les Phrases et les Chiffres, le disque 2 les Nombres, et les disques 3 et 4 les Mots (Logatomes CVC).

V. CONCLUSION

La base de données BDBRUIT est actuellement disponible sur CD-ROM (standard ISO-9660) auprès du GDR-PRC "CHM". La documentation sur cette base de données ainsi que sur les autres bases du GDR-PRC, sera mise en ligne sur un serveur FTP et pourra être feuilletée via le réseau informatique en utilisant le logiciel diffusant des documents hypermedia: NCSA "Mosaic" [13]. Une maquette réaliste a déjà été développée.

10 locuteurs	CVC	PHRASES	NOMBRES
Chaque locuteur a prononcé:	3 X 82	30	5 X 30
chaque item a été prononcé au total:	30 fois	10 fois	50 fois

fig 3 : Répétitions et occurrences

D'autre part, le logiciel GERSONS [14] développé pour la gestion de la base de données des sons du français (BDSONS) est adapté pour permettre la gestion de cette base. Ce logiciel, doté d'une interface utilisateur conviviale, facilite l'exploitation de la base (sélection d'items, sélection de locuteurs, sélection de conditions d'enregistrement, extraction de sons...). Il est opérationnel sur station de type compatible PC.

Dans la mesure où les données linguistiques de BDBRUIT se trouvent prononcées par d'autres locuteurs dans EUROM1 (partie française), on pourra disposer de références supplémentaires qui permettront d'étendre les résultats obtenus pour l'effet Lombard. La disponibilité d'enregistrements en cours de réalisation pour un certain nombre de langues d'Europe (EUROM1, autres langues [15]) rendra possible une généralisation de ces mêmes résultats grâce à des comparaisons multilingues. S'il apparaissait que les faits recensés ne relèvent pas d'une langue particulière mais se retrouvent dans des contextes comparables dans d'autres langues, on pourrait s'orienter vers des interprétations de type physiologique au sens large (stratégies générales de la communication linguistique dans le bruit).

Bibliographie:

[1] Lombard E. (1911)
Le signe de l'élévation de la voix
Ann. Maladies oreille, larynx, nez, pharynx, 37,101-119

[2] Dejonckere P.H., Pepin F. (1963)
Etude de l'effet Lombard par la mesure du niveau sonore équivalent.
Folia Phonatrica 35, 310-315

[3] Anglade Y., Junqua J.C. (1990)
Etude acoustique de l'effet Lombard sur des phonèmes de l'anglais américain dans le cadre de mots isolés.
18èmes J.E.P., SFA, Montréal (Canada), pp.138-142.

[4] Descout R., Sérignat J-F., Cervantes O., Carré R. (1986).
BDSONS: une base de données des sons du français.
12th Internat. Congress on Acoustics, A4-7, july Toronto (Canada).

[5] Zeiliger J., Sérignat J-F., Autesserre D., Dolmazon J-M (1992).
EUROM1: une base de données "Parole" multilingue Partie française.
19èmes J.E.P., SFA, pp 303-306, 19-22 mai, Bruxelles (Belgique).

[6] Dolmazon J-M., Benoit C., Gauvain J-L., Perennou G. (1987).
Le projet européen SAM : évaluation multilingue des dispositifs d'entrée-sortie vocales.
16èmes J.E.P., SFA, pp 34-317, octobre, Hammamet (Tunisie).

[7] Zeiliger J., Sérignat J-F.(1991)
"EUROPEC software , User's Guide Release 4.1"
SAM-ICP-045, March, Source: Institut de la Communication Parlée, Grenoble.

[8] Combescure P. (1979/1980)
Phrases phonétiquement équilibrées.
Rapport Recherches /Acoustique, volume VI, p 45-60, CNET, Lannion.

[9] Sérignat J.F., Zeiliger J. (1992)
"Annex ICP 1 - List of EUROM1 database corpora (French Part)".
SAM- ICP-051, January, Source: Institut de la Communication Parlée, Grenoble.

[10] Carton F., Rossi M., Autesserre D., Leon P. (1983)
Les accents des français.
Collection "de bouche à oreille", Hachette, Paris.

[11] Tomlinson M. (1990)
Guide to database generation - Recording protocols.
SAM-RSRE-013, April 3rd. Source: RSRE, Great Malvern (UK).

[12] Autesserre D., Nishinuma Y., Delran-Bado S. (1992).
Respiration et phonation en situation de dialogue oral.
Actes du séminaire Dialogue, CNRS/LIMSI, pp.19-30.

[13] NCSA Mosaic version 2.0 (1993)
National Center for Supercomputing Applications.
University of Illinois, Urbana-Champaign, (USA).

[14] Sérignat J-F., Dolmazon J-M. (1993).
BDSONS- GERSONS: système multi-base et multilingue de gestion de sons de parole.
Forum des recherches en Informatique, pp 84-85, 2-3 juin, Ecole Polytechnique, Palaiseau.

[15] Sherwood T., Fuller H., (1992)
Guide to EUROM1 speech database
SAM-NPL-102, April 21st, Source: NPL, Teddington, UK.

TEST ET MISE-AU-POINT DES PAQUETS DE REGLES DE PRODUCTION POUR LA COMMUNICATION PARLEE

P. MUNTEANU ; J.F. SERIGNAT

INSTITUT DE LA COMMUNICATION PARLEE (ICP)
46, avenue Félix Viallet 38031 GRENOBLE CEDEX

Résumé

Cet article présente le contexte et les objectifs du développement d'un système permettant le test et la mise au point de paquets de règles de production pour tout type d'étiquetage dans le domaine de la parole. Ce travail permet d'envisager efficacement la validation et l'optimisation de systèmes experts à règles de production pour la reconnaissance automatique de la parole, indépendamment du contexte informatique ou logiciel de développement de ces systèmes. La méthodologie, pour la modélisation et la mise en œuvre dans un environnement orienté objet, est décrite pour le paradigme de l'étiquetage prosodique.

Mots-clés: Base de données objets, base de connaissances, système expert, base de règles, règles de production, objets parole, prosodie

I. INTRODUCTION

Les travaux qui sont à la base de cette communication s'inscrivent dans le cadre du projet SIDOC-Parole (Système d'Intégration de Données et de Connaissances sur la Parole), développé à l'Institut de la Communication Parlée de Grenoble, qui "vise à fournir un environnement de gestion d'informations sur la parole pour appuyer la recherche scientifique".

Ce travail se positionne clairement dans l'axe du développement et de l'optimisation de systèmes experts à règles de production (donc "basés connaissances") pour la reconnaissance automatique de la parole, le D.A.P., et plus généralement tout type d'étiquetage du signal acoustique, correspondant à différents niveaux de représentation de la connaissance : phonétique, prosodique, articulatoire, syntaxique, sémantique, etc... En effet, les connaissances et l'expérience acquises par les experts dans ces domaines ne sont pas encore consensuelles et évoluent continuellement. En conséquence, un formalisme déclaratif, simple à modifier et avec des potentialités d'explication est préféré aux langages de programmation classiques (algorithmiques).

Cet article présente la démarche que nous avons suivie pour intégrer dans le SIDOC des fonctionnalités de gestion, test et mise-au-point des paquets de règles de production en décrivant un exemple concret pour la prosodie.

II. LE SIDOC ET SON CONTEXTE

Le SIDOC est un vaste projet initié à l'ICP depuis plusieurs années [1], [2], [3], dont l'objectif principal est le développement d'une plateforme multimédia pour l'apprentissage et l'expertise des connaissances dans le domaine de la parole. Compte tenu de l'impressionnante complexité du domaine, nous avons choisi un modèle orienté objet pour la représentation des données et des connaissances en parole et une première architecture de base de données et de connaissances a été définie [4].

Un prototype du système a été développé sur une station de travail de type P.C., spécialement équipée pour la gestion et l'écoute des éléments de signal vocal enregistré. Il offre les fonctionnalités d'un système de gestion de base de données orientée objet pour la BDSONS (Base de Données des Sons du français) [5] et pour les descripteurs associés. Des facilités d'accès vers les systèmes externes d'édition et de traitement du signal, de même que la possibilité d'importation / exportation de données ont été ajoutées (fig. 1.) [6].

Une des principales difficultés que rencontrent les chercheurs dans le domaine de la reconnaissance et de la synthèse de parole est l'immense variabilité, inter- et même intra-locuteur, du signal de la parole. D'où la nécessité de disposer de larges corpus de parole enregistrés par de nombreux locuteurs, ce qui a poussé, depuis cinq à dix ans, les divers laboratoires ou organismes de recherche à fournir un effort considérable pour l'enregistrement de larges bases de données [7], [8], [9], [10].

En conséquence, il serait utile de pouvoir disposer d'un environnement intégré permettant de tester d'une manière systématique sur des volumes importants de données, des ensembles ("paquets") de règles de production (forme : "SI condition ALORS action") développées pour les applications d'étiquetage en vue de la reconnaissance. L'objectif principal du SIDOC est de

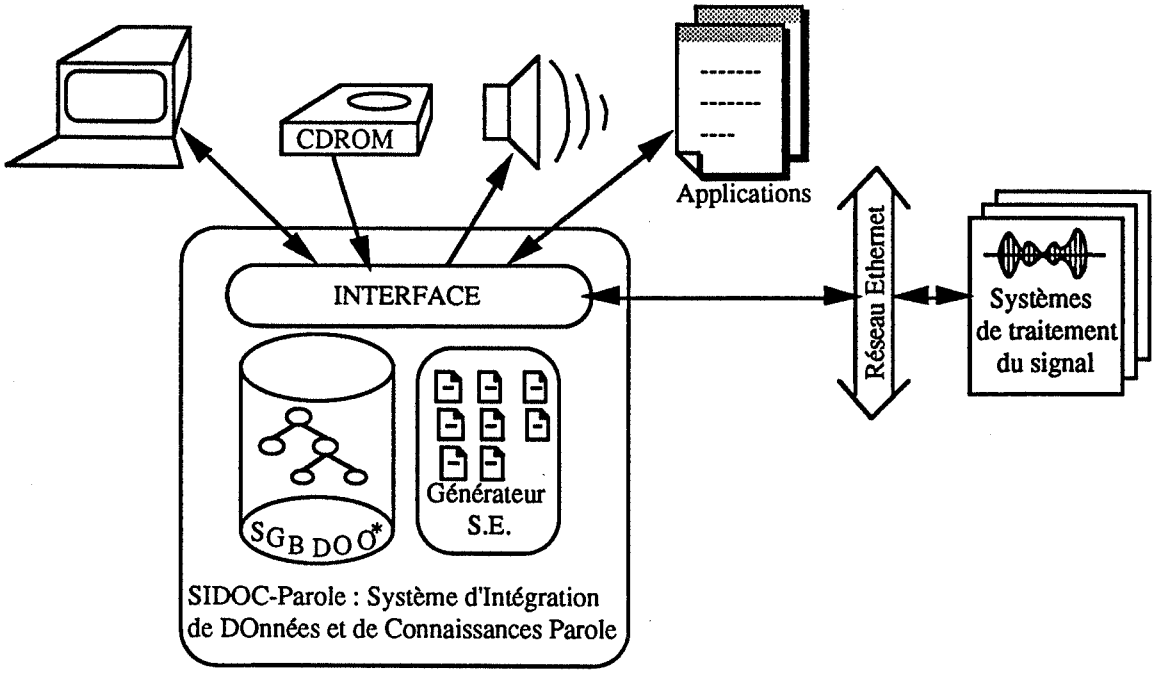


fig. 1. Synoptique des fonctionnalités du SIDOC
* SGBDOO : Système de Gestion de Bases de Données Orienté Objet.

tirer avantage des bases de données standards (comme BDSONS) pour proposer des critères d'appréciation objective ("benchmarks") afin d'améliorer les performances des systèmes de reconnaissance ou de comparer entre elles les réalisations similaires.

III. LE PARADIGME DE L'ETIQUETAGE PROSODIQUE

Pour valider ces nouvelles fonctionnalités du SIDOC dans une situation concrète et représentative d'un domaine, nous avons utilisé un paquet de règles en prosodie développé à l'I.C.P. à des fins de reconnaissance de parole [11], [12]. Au cours de cet article nous allons nous rapporter à cet exemple pour fixer les idées.

Le but de ce paquet de règles est de détecter sur le signal vocal des traits prosodiques (début / fin de mot lexical, mot grammatical, début / fin de phrase, groupes majeurs ou mineurs) et de fournir ces indications aux autres étages de traitement qui interviennent dans la reconnaissance de la parole.

Le paquet est destiné à s'exécuter sur chaque noyau vocalique préalablement détecté dans le signal à l'aide de critères sur l'énergie et la fréquence fondamentale. Les prémisses des règles testent différents paramètres du noyau vocalique courant et de ses voisins : durée, énergie, fréquence fondamentale et pause avant noyau. Une quantification en 4 niveaux est en plus requise pour certains de ces paramètres.

Voici un exemple de règle du paquet, telle que rédigée par l'expert :

règle FML1 (fin de mot lexical)
SI D(n)-D(n-1)≥D(n-1)·1,5
ET F0(n)>ligne de régression de F0
ET F0=4
ET E≥3
ALORS FML

où : D=durée effective du signal entre le noyau vocalique courant et le noyau vocalique précédant, F0=fréquence fondamentale, E=énergie, n=noyau vocalique courant et (n-1)=noyau vocalique précédent.

La figure 3. présente un synoptique de l'enchaînement des opérations mises en jeu dans le paradigme de l'étiquetage prosodique (P.E.P.) que l'on explicitera au paragraphe VI.

IV. PLACE DU SIDOC DANS LE DOMAINE DE LA VALIDATION DES SYSTEMES EXPERTS

Nos travaux s'insèrent dans le domaine de la validation des systèmes experts. Ce domaine, nouveau et complexe, présente certaines particularités par rapport à la validation des systèmes informatiques classiques. La littérature qui traite de ce problème [13], [14] fait généralement une distinction entre les différents aspects de la validation : la vérification, l'évaluation et le test. Par ailleurs, la validation peut s'effectuer sur la base de connaissances, sur le moteur d'inférences ou sur le système expert dans son ensemble.

L'efficacité du test dans la validation des systèmes experts est dépendante de la représentativité de

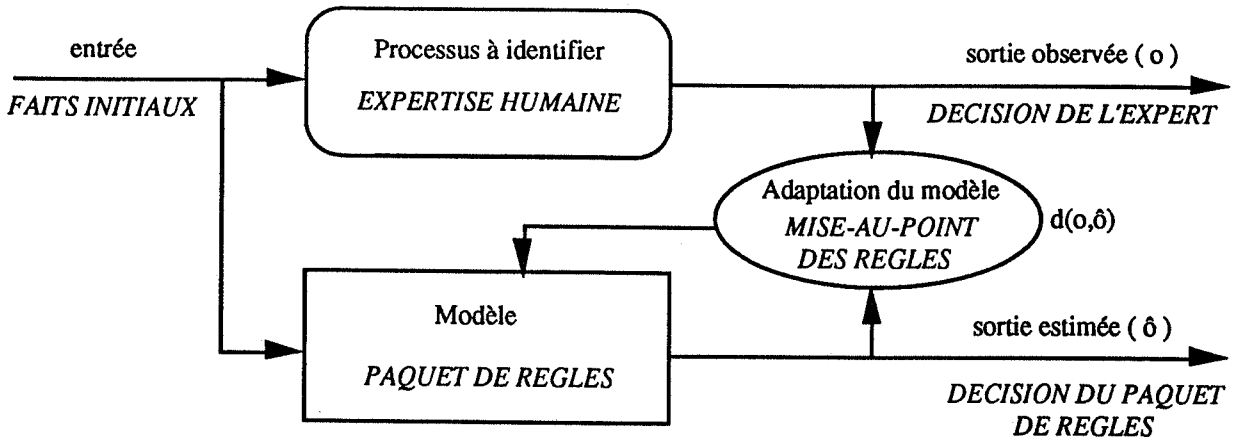


fig. 2. Mise-au-point des paquets de règles vue comme identification de processus

l'ensemble des contextes d'exécution choisis. Certaines heuristiques ont été proposées dans la littérature pour réaliser ce choix [13].

Dans cet article nous présentons les outils que nous avons développés pour le *test* des paquets de règles sur des ensembles de données (espaces) préalablement choisies par l'utilisateur expert, en fonction de divers critères. Pour la sélection de ces données et leur regroupement dans des espaces, en réponse aux requêtes de l'utilisateur, le système est pourvu d'interfaces spécifiques.

V. LA SEMANTIQUE DU SUCCES

Les résultats des tests sont d'habitude fournis sous forme du pourcentage des succès (ou des échecs) enregistrés sur l'ensemble des situations de test. Mais la sémantique des notions telles que "Succès" et "Echec" est très dépendante des buts poursuivis par le développeur du paquet de règles. Ces buts peuvent varier, même dans le cadre d'une seule application, en fonction de l'étape de développement où se trouve le paquet de règles. Par exemple, dans une première étape on pourrait vouloir tester si un certain paquet de règles est complet, c'est à dire s'il fournit une réponse (n'importe laquelle, pour le moment) dans toutes les situations. Ensuite, pour les règles prosodiques par exemple, il pourrait être intéressant de vérifier que le paquet ne fait pas de fausse détection de trait, plus "coûteuse" qu'une omission de détection. Ultérieurement, dans une phase plus avancée, l'obtention de tous les marqueurs prosodiques pourrait devenir l'objectif du test.

Il devient donc évident que l'objectif du test doit être fourni explicitement par l'utilisateur, de la même façon que l'espace sur lequel le système doit faire tourner le paquet de règles. Les spécifications qui peuvent être testées d'une manière automatique doivent donc être formalisables, sous forme de prédicats, par exemple (ce qui élimine les spécifications peu ou pas du tout formalisables, assez fréquentes dans le domaine des systèmes experts).

D'habitude, la qualité des résultats fournis par un système expert est évaluée par rapport aux résultats obtenus par les experts humains placés dans des conditions similaires. C'est le cas des systèmes experts en médecine, par exemple, où les bases de données de diagnostiques, archivées par les hôpitaux, constituent les références pour les tests. Ainsi, pour ce type de test il faut avoir accès à des volumes importants de résultats d'expertise, afin de pouvoir valoriser une base de données comme celle gérée par le SIDOC (constituée principalement au départ des "faits" initiaux, de même nature que ceux fournis aux systèmes experts).

L'obtention de ces références par d'autres moyens que l'expertise (des traitements automatiques, par exemple) n'est généralement pas possible. En effet, les problèmes, pour lesquels on développe des systèmes experts, n'ont pas (ou pas encore) de solutions algorithmiques (ceci est un des motifs pour lesquels on utilise les systèmes experts). Le travail d'étiquetage par un Expert des bases de données standards, comme BDSONS, même s'il est laborieux et prend du temps, est donc absolument nécessaire pour servir de référence dans le test des systèmes experts.

Au niveau méthodologique, on peut proposer une analogie pour mieux fixer les idées (fig. 2.) : si on regarde la mise-au-point des paquets de règles comme une modélisation (identification) incrémentale du processus d'expertise humaine, il est absolument nécessaire de disposer de sorties réelles (décisions de l'expert) et d'une distance (prédicat à vérifier) pour apprécier les performances du modèle (paquet de règles).

VI. PRESENTATION DU SYSTEME

Dans ce paragraphe nous allons présenter les fonctionnalités du système liées à la gestion et au test des paquets de règles de production en suivant le scénario général de ces opérations illustré par le paradigme de

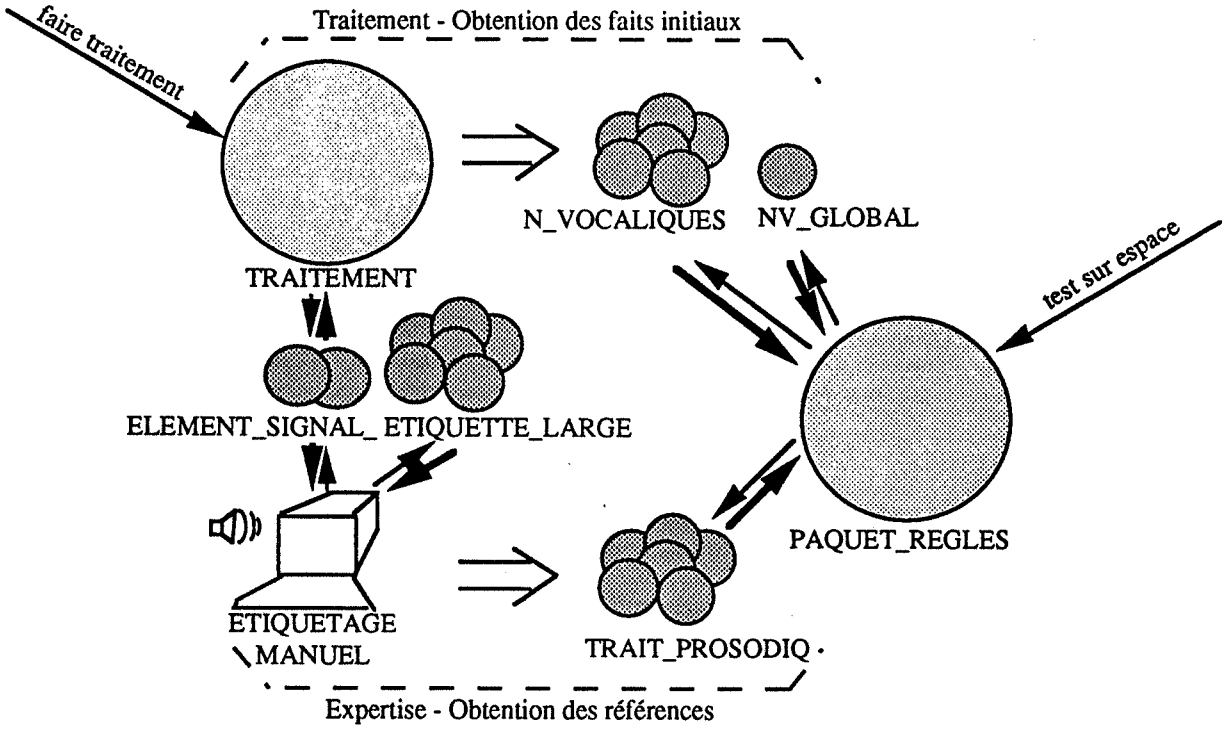


fig. 3. Intégration et test des règles prosodiques dans le SIDOC

l'étiquetage prosodique (P.E.P.) (fig. 3.). On considère connus les concepts de base des modèles à objets.

Pour intégrer et valider un paquet de règles dans le SIDOC, l'utilisateur doit parcourir les étapes suivantes :

1. *Identification (et obtention) des objets concernés par le paquet de règles (Espace d'observation)*

Les prémisses des règles du paquet à intégrer testent généralement les valeurs de certains paramètres. Si ces données ne sont pas gérées par le système au moment de l'intégration, les fonctionnalités de création / modification d'objets génériques du SIDOC peuvent être utilisées pour modéliser ces nouveaux aspects du domaine. Les instances correspondantes à l'espace de test peuvent être créées à l'aide de l'interface vers les systèmes coopérateurs d'édition et de traitement du signal et d'exportation / importation de données.

P.E.P. : Les noyaux vocaliques et leurs paramètres n'existaient pas dans le SIDOC. Nous avons donc dû faire tourner un traitement de détection de noyaux vocaliques et de calcul de paramètres [15]) sur les éléments de signal sélectionnés pour le test. Pour accueillir les résultats de ce traitement nous avons créé deux nouveaux objets génériques, spécialisations du RESULTAT_TRAITEMENT: NOYAU_VOCALIQUE et NV_GLOBAL.

2. *Identification (et obtention) des références d'expertise*

Comme nous l'avons déjà mentionné dans le paragraphe précédent, pour juger les performances d'un paquet de règles nous devons disposer d'un volume important de résultats d'expertise humaine. Certaines informations de ce type (comme, par exemple, les étiquettes phonétiques pour les éléments de signal de BDSONS) sont ajoutées dans le système au fur et à mesure que le travail d'étiquetage avance. Les mêmes fonctionnalités de création / mofification d'objets génériques et d'instances servent à la modélisation et à l'intégration de ces résultats.

P.E.P. : L'expert en prosodie qui est l'auteur du paquet de règles a réalisé un étiquetage manuel des éléments de signal choisis en indiquant la correspondance entre les traits prosodiques et les étiquettes phonétiques déjà existantes dans le SIDOC. Nous avons intégré les résultats de cet étiquetage comme instances d'un nouvel objet générique, spécialement créé: TRAIT_PROSODIQ.

3. *Intégration effective du paquet de règles*

Les paquets de règles sont représentés dans le SIDOC sous forme d'objets. Cette variante de modélisation a été choisie autant par souci d'homogénéité avec le formalisme utilisé pour la représentation des données que pour profiter des mécanismes de gestion existants.

Il faut souligner une particularité importante du SIDOC par rapport à d'autres systèmes de représentation des connaissances en ce qui concerne les règles de production : même si cette forme de représentation peut être utilisée, après toutes les vérifications nécessaires,

d'une façon similaire aux traitements, pour produire de nouvelles connaissances dans le SIDOC, elle est conçue principalement pour assurer le cadre de validation des paquets de règles qui seront ensuite utilisés dans des systèmes experts, autres que SIDOC.

Cette particularité nous à imposé une contrainte d'implémentation : *les règles doivent être le plus indépendantes possible du contexte d'utilisation.* Dans la structure d'un paquet de règles représenté dans le SIDOC, nous nous sommes attachés à délimiter clairement les "connaissances de type expert" à vérifier, de ce qui constitue l'interface d'entrée (l'obtention des valeurs pour les variables utilisées dans les règles) ou de sortie (l'éventuelle intégration dans le système des connaissances produites). L'effort de programmation requis pour la conversion d'un paquet de règles dans le format de SIDOC est donc concentré sur l'aspect "communication", les règles elles-mêmes n'étant pas directement affectées.

P.E.P. : Le paquet de règles prosodiques a été intégré dans le SIDOC comme instance de l'objet générique PAQUET_REGLES. La "traduction" des règles, fournies dans un pseudo-code proche du langage naturel, dans le langage de SIDOC a été nécessaire mais, apparemment, cette opération ne doit pas poser de problème, même pour les non-spécialistes en programmation. L'obtention des valeurs des paramètres testés par les règles s'effectue par envoi de messages vers les instances correspondantes des objets génériques NOYAU_VOCALIQUE et NV_GLOBAL.

4. *Test et mise au point du paquet de règles*

La "boite à outils" du SIDOC, pour la mise au point des paquets de règles, est constituée par la liste des méthodes attachées à l'objet générique PAQUET_REGLES. En conséquence on va y trouver des "méthodes" pour :

- éditer les paquets de règles (*voir / modifier*);
- faire tourner le paquet de règles sur toutes les instances sélectionnées dans un espace et calculer le taux de réussite (*test sur espace*);
- déclencher la consultation du paquet pour une seule instance en chaînage arrière, chaînage avant ou pour tirer une certaine règle (*consultation, chaînage avant, tirer une règle*);
- fournir des explications sur le tirage d'une règle, l'obtention de la valeur d'une variable ou sur la consultation dans son ensemble (*pourquoi?, comment?, justification*);
- présenter l'interdépendance des règles sous une forme graphique suggestive (*arbre*).

Toutes ces "méthodes" ont été développées en exploitant les fonctionnalités du générateur de système expert GURU [16], qui est à la base du développement du SIDOC.

Pour utiliser ces outils pendant la mise au point des paquets de règles, nous proposons un "mode d'emploi" basé sur l'itération du cycle d'opérations suivant :

1. sélection dans un espace des instances qui présentent des propriétés intéressantes pour l'étude d'un certain aspect du fonctionnement du paquet de règles. Cette sélection peut être faite grâce à une interface spécialisée, implémentée comme méthode de création d'instances de l'objet générique ESPACE. Cette interface propose à l'utilisateur la possibilité de sélectionner les instances une à une ou d'adresser des requêtes de type SQL (Structured Query Language) au système.

2. test du paquet de règles sur l'espace respectif par l'intermédiaire de la méthode *Test sur espace.* Un prédicat à vérifier, qui formalise les attentes de l'utilisateur en ce qui concerne les résultats obtenus par le paquet de règles, est demandé avant le test. Une demande de *Justification*, consécutive au test, fournit le résultat de la vérification de ce prédicat pour chacune des instances : SUCCES ou ECHEC.

3. focalisation sur les cas d'ECHEC : *Consultation* du paquet de règles pour ces cas et demande de *Justification* détaillée de l'insuccès. D'autres méthodes (*Chaînage avant, Tirer une règle, Pourquoi ?* ou *Comment ?*) peuvent aussi être utilisées pour déterminer les causes de l'insuccès.

4. ajustement en conséquence des règles (seuils, coefficients, etc.).

Pour simplifier la détection des causes d'insuccès, une décomposition préalable du paquet de règles en des sous-paquets est souhaitable. Ces sous-paquets, testés individuellement, seront ensuite regroupés dans des systèmes plus complexes pour étudier leur comportement d'ensemble.

P.E.P. : Les résultats peu satisfaisants d'un premier test, effectué sur 2 éléments de signal traités (175 noyaux vocaliques pour un et 168 pour l'autre), nous ont suggérés l'existence d'une erreur systématique, ce qui nous a permis de découvrir des incohérences dans les résultats de traitement. Après avoir corrigé ces erreurs nous sommes passés à une décomposition du paquet de règles par classes des marqueurs prosodiques positionnés, que nous avons testées ensuite suivant la méthodologie décrite en haut. L'itération de ces tests nous a permis une meilleure calibration des seuils et coefficients, ce qui a eu comme effet une amélioration du score du paquet de règles.

Le parcours de toutes ces étapes, pour une situation concrète "le paquet de règles prosodiques" nous a permis de valider les nouvelles fonctionnalités intégrées dans le SIDOC et l'enchaînement des opérations nécessaires

pendant leur utilisation. Les chercheurs qui travaillent sur la mise au point de ce paquet de règles se sont déclarés intéressés par ces fonctionnalités et ont décidé d'utiliser le SIDOC pour continuer leurs tests.

VII. CONCLUSION ET PERSPECTIVES

Dans l'état actuel, le système offre un cadre unitaire et ergonomique pour le test des connaissances explicitées par les experts sous forme de règles de production. Ce travail constitue une avancée sur le chemin qui mène des données vers les connaissances, que s'est proposé de parcourir le SIDOC depuis son origine.

Les résultats présentés dans cet article s'inscrivent dans le cadre des recherches sur la validation et l'intégration des connaissances apportées par les experts humains. Mais les travaux menés à plus long terme dans le cadre du SIDOC vont aussi dans le sens de l'exploitation de l'important volume de données gérées, pour l'apprentissage et la production de connaissances par le système lui-même.

BIBLIOGRAPHIE

[1] O. Cervantes (1988)
Bases de données et d'objets multi-média pour la recherche sur la parole.
Thèse de Docteur en Informatique INPG, janvier, GRENOBLE.

[2] Y. Fernandez (1989)
Gestion de connaissances pour des applications du domaine de la parole.
Thèse de Doctorat N.R., Spécialité Informatique, INPG, décembre, GRENOBLE.

[3] H. Belbachir (1991)
Maintien de la cohérence et interface utilisateur dans une base orientée objet. Application à la parole.
Thèse de Doctorat es Sciences Informatique, préparée à l'ICP-Grenoble, Université d'Oran Es Senia (Algérie).

[4] O. Cervantes, J.F. Serignat (1988)
Un modèle centré objet pour la base de données et de connaissances parole.
Journal d'Acoustique, SFA, T1, N°4, pp 309-315, décembre.

[5] A. Neagu (1993)
Optimisation d'un gestionnaire de données et de connaissances sur la parole.
Mémoire de diplôme d'Ingénieur Inst.Polytech. Bucarest, rapport interne ICP, GRENOBLE.

[6] R.Descout, J.F. Serignat, O.Cervantes, R.Carré (1986)
BDSONS : Une base de données des sons du français
12-th Internal. Congress. on Acoustics, A 4-7, july TORONTO (Canada).

[7] P. Price, W.M. Fisher, J. Bernstein, D.S.Pallet (1988)
The DARPA 1000 word Ressource Management Database for Continuous Speech Recognition
Proc. ICASSP, vol.2, pp 651-654, April, NEW YORK (USA).

[8] L.F. Lamel, J.L. Gauvain, M. Eskenazi (1991)
BREF, a Large Vocabulary Spoken Corpus for French
EUROSPEECH 91 : 2-nd European Conf. on Speech Com. and Technology, vol.2, pp 505-508, September, GENOVA (Italy).

[9] V. Zue, J. Glass, D. Goodine, L. Hirshman, H. Leung, M. Philips, J. Polifroni, S. Seneff (1991)
The MIT ATIS System : Preliminary Development, Spontaneus Speech Data Collection and Performance Evaluation.
EUROSPEECH 91 : 2-nd European Conf. on Speech Com. and Technology, vol 2, pp 537-540, September, GENOVA (Italy).

[10] J. Zeiliger, J.F. Serignat, D. Autesserre, J.M. Dolmazon (1992)
EUROM1 : Une base de données "Parole" multilingue. Partie française.
19-emes J.E.P., SFA, pp 303-306, Mai, BRUXELLES (Belgique).

[11] M.K. Nasri, G. Caelen-Haumont, J. Caelen (1989)
Using Prosodic Rules in Speech Recognition Expert System.
ICASSP Proceedings, IEEE, vol 1, Speech Processing 1, pp 671-674, GLASGOW (Scotland).

[12] M.K. Nasri, G. Caelen-Haumont, J. Caelen (1990)
Utilisation de règles prosodiques en reconnaissance de la parole.
Actes des 18-èmes Journées d'Etudes sur la Parole, SFA CNRS, MONTREAL (Canada).

[13] P. Meseguer (1992)
Towards a Conceptual Framework for Expert System Validation.
AICOM, vol.5, no.3, september.

[14] O.J. Mengshoel (1993)
Knowledge Validation : Principles and Practice.
IEEE EXPERT, vol 8, no 3, pp 62-68, June

[15] B. Caillaud, J.-Y. Antoine, J. Caelen, G. Caelen-Haumont (1994)
Micro : Un système multi-agents pour la compréhension de la parole.
9-ème Congrès RFIA, AFCET, janvier, PARIS.

[16] Micro Data Base Systems, Inc (M.D.B.S.) (1990)
GURU, générateur de Systèmes experts, ver 2.0.
Manuel de Référence Guru, Traduct. ISE Cegos, BOULOGNE-BILLANCOURT.

LEMMATISATION DU DICTIONNAIRE PHONÉTIQUE DU FRANÇAIS PARLÉ « DE À à ZUT »

Danièle DUJARDIN[1], Rabia BELRHALI[2], Louis-Jean BOË[2], Jacques COURTIN[1]

1 Équipe TRILAN, Laboratoire de Génie Informatique
IMAG Campus BP 53, 38041 Grenoble cedex 9
daniele.dujardin@imag.fr, jacques.courtin@imag.fr

2 INSTITUT DE LA COMMUNICATION PARLÉE
CNRS URA 368, INPG/Univ. Stendhal, BP 25 38040 Grenoble cedex 9
belrhali@icp.grenet.fr, boe@icp.grenet.fr

Résumé

L'objectif de cette étude vise à la lemmatisation du lexique du français parlé « De À à Zut ». Ce lexique de formes orthographiques du corpus phonétique BDPHO auxquelles sont associées le nombre d'occurrences et les cohortes. Pour cela, nous avons utilisé un système de traitement de langue naturelle PILAF qui a permis l'analyse morphologique et la lemmatisation de ces formes. La version lemmatisée de BDPHO regroupe, pour chaque forme fléchie, les informations des différents niveaux linguistiques : lexical, phonétique, morphologique, syntaxique. La présence des liens entre les contextes orthographiques et phonétiques et la forme fléchie nous permet de proposer une nouvelle méthode de travail dans le traitement des corpus oraux.

Mots clés : *Phonétique, corpus, lemmatisation, base de données, communication parlée, morphologie, TALN.*

I. INTRODUCTION

Le but de cette étude est la lemmatisation d'une base de données lexicale orthographique-phonétique du français parlé : BDPHO [BOË 92a]. Cette base de données a été élaborée par l'École Nationale Supérieure des Télécommunications (département Signal) et l'Institut de la Communication Parlée de Grenoble à partir d'un corpus de parole enregistrée d'environ 10 heures, noté phonétiquement par des experts (une trentaine de locuteurs, plus de 300 000 sons et 102 000 mots). La restitution de la forme orthographique a permis de constituer BDPHO sur la base de 7 590 entrées orthographiques (pour 7 221 formes phonétiques, dont 1 386 variantes). Elle comporte six champs : l'entrée orthographique, le nombre d'occurrences total, la (ou les) forme(s) phonétique(s), la cohorte correspondante, le nombre d'occurrences et une indexation permettant la localisation des 10 premières occurrences dans le corpus. BDPHO a été développée avec un matériel Macintosh, dans un environnement HyperCard (langage HyperTalk); elle se compose de trois parties de corpus dans leurs formes orthographique et phonétique et deux bases orthographiques-phonétiques lexicales.

Un dictionnaire « De À à Zut » [BOË 92b], a été élaboré à partir de cette base de données. C'est un dictionnaire de formes auxquelles sont associées le nombre d'occurrences relevées dans le corpus, les variantes de prononciation et la structure du mot phonétique sous forme de cohorte. Le dictionnaire a été publié aux éditions ELLUG en 1992.

La lemmatisation de ce dictionnaire se situe dans le cadre du GDR Programme de recherche concertée « Communication homme-machine » Pôles Parole & Langue Naturelle et fait partie du projet ORALL (Organisation et Accès à de Larges Lexiques en vue du traitement de la Parole) maintenant s'intègre dans le Groupe de Travail (GT5) Lexique du PRC Informatique.

Le système PILAF utilisé pour la lemmatisation a été mis au point par l'équipe TRILAN. Dans cet article, nous présenterons les différents outils nécessaires à cette étude, la méthodologie que nous avons suivie, et enfin nous ferrons une première lecture du traitement.

II. LES ÉLÉMENTS EN PRÉSENCE

1. Les corpus

1.1. Origine et description des corpus

BDPHO, constitué de trois parties, représente précisément 304 752 sons.

• Le corpus n° 1 (86 360 sons) a été constitué à l'Institut de Phonétique de Grenoble sous la direction de R. GSELL. Il est divisé en cinq parties (au total 2 heures 1/2) enregistrées entre 1967 et 1968 : 1/2 heure d'informations de France-Inter, et 2 heures de conférences de France Culture : « Inter-Opinions » (à propos du livre 60 millions de Français), entretien MALRAUX-BOISDEFFRE, conférence sur MAUPASSANT, « Recherche de Notre Temps » et « La province aujourd'hui ». Cet ensemble est donc représentatif d'un langage parlé cultivé et assez formel. La transcription est fine et comporte de nombreux diacritiques. Elle a été assurée, comme la saisie sur support informatique, par G. RAKOTOFIRINGA et C. GERMI.

• Le corpus n° 2, avec ses 201 281 sons, est le plus important ; il a été constitué dans le cadre d'un projet à l'Université de Californie (Santa Barbara) par A. MALÉCOT en 1972. Nous avons pu en disposer, avec son autorisation, grâce à l'obligeance et l'efficacité de G. CHOLLET. Ce corpus est constitué de cinquante conversations informelles d'une demi-heure, sur des sujets très divers, avec des membres de « l'intelligentsia » parisienne (professeurs, avocats, médecins, artistes etc.). Dans la version originale du corpus, il y avait, outre la transcription phonétique, que

nous avons exclusivement conservée, d'autres informations, prosodiques et grammaticales. Il s'agit d'un niveau de langue un peu plus familier que celui du premier corpus. La transcription est plutôt large. La version originale du corpus comportait des informations équivalentes à des diacritiques, mais nous n'avons pas pu les récupérer. Nous avons remplacé par un signe non ambigu phonétiquement (x), les noms de personne et un certain nombre de mots ou expressions de langue anglaise ou américaine qui avaient été transcrits comme du français.

• Le corpus n° 3 (17 111 sons) a été recueilli à l'Institut de la Communication Parlée par J. VAN EIBERGEN [VAN EIBERGEN 85]. Il comporte les transcriptions de 16 brèves conversations, par 16 locuteurs, d'origine linguistique et socio-professionnelle variées (4 adolescents de moins de 20 ans, 8 adultes de 20 à 60 ans, 4 adultes de 60 à 80 ans). Il s'agit de conversations, en situation informelle, entre familiers. Le niveau de langue est celui d'un français très spontané. La transcription est très proche de celle utilisée pour le corpus n° 1. Une attention particulière a été portée par l'auteur pour la notation du schwa.

1.2. Codification

Étant données leurs origines différentes, ces corpus n'étaient pas codés de la même façon. Nous avons défini une représentation du sous-ensemble de l'API utilisé pour le français, avec quelques caractères additionnels. Les différences entre les corpus sont, en ce qui concerne :

• Les sons : les différences portent sur les deux a et sur le schwa.

Corpus n° 1 : [a] et [ɑ] sont différenciés
Corpus n° 2 : un seul /a/, noté [a]
Corpus n° 3 : [a] et [ɑ] sont distingués

comme dans le corpus n° 1. Quatre symboles différents avaient été utilisés pour noter différentes réalisations du schwa : non réalisation, réalisation [ə], « euh » d'hésitation réalisé [ə], « euh » d'hésitation non linguistique. Nous avons supprimé la marque de non réalisation et confondu les autres réalisations en [ə].

• Les séparateurs : le corpus 2 ne comporte pas de séparateurs de groupes rythmiques (/), mais seulement des séparateurs de phrases (#). Les liaisons sont transcrites mais rattachées au mot précédent sans indication particulière. Les deux autres, corpus 1 et corpus 3 comportent des séparateurs de groupes rythmiques et de phrases distincts, et les liaisons sont identifiées en tant que telles par le symbole (–) rattaché au mot générateur.

• Les diacritiques : le corpus 2 ne comporte pas de diacritiques, tandis que les corpus 1 et 3 possèdent des indications d'assourdissement (̥) et de sonorisation (̬), d'accentuation simple (') et d'insistance ("), ainsi que d'allongement vocalique (:).

1.3. Contenu du corpus

Au total l'ensemble du corpus comporte 102 137 occurrences lexicales, soit 7 221 formes phonétiques différentes. La combinatoire des « polysons » est loin d'être épuisée : si les « bisons » apparaissent très nombreux (87% des possibilités), les « trisons » et surtout les « quadrisons » sont relativement moins représentés (respectivement 30% et 3.8%) ; de la liste émergent déjà des mots ou des suites de mots (alors, parce que, il y a, de la, avec, et par, crois, voir...). Les liaisons qui engendrent les sons [d, n, p, R, t, z] représentent environ 6% des occurrences de mots du corpus et sont très inégalement réparties, trois d'entre elles [n, t, z] totalisant plus de 90% de l'ensemble. La mise en regard des 50 premières occurrences du corpus avec celles publiées par G. ENGWALL [ENGWALL 84] (Vocabulaire du Roman français : 500 096 occurrences) et celles des Listes Orthographiques de Base (LOB0) de N. CATACH [CATACH 84] fait apparaître de très nettes ressemblances (plus de 30 formes ou entrées communes) et l'émergence de lemmes caractéristiques de l'oral (ça, y, alors, très, oui, enfin, parce que, moi, quand, puis, euh !), avec de sérieuses différences quantitatives en faveur de l'oral (pour est, c', pas, on, ça, ce, y, pour, bien, alors, très, oui, enfin, parce que, fait, si, même, là, euh) et très rarement en sa défaveur (négation ne souvent omise). Au niveau des cohortes 25% d'entre elles couvrent 90% des possibilités et 44 permettent d'atteindre 95% du total.

2. La base de données

Après restitution orthographique, le lexique a été constitué sur la base de 7 000 entrées orthographiques, compte tenu des homophones hétérographes. Nous avons relevé 1 386 variantes qui seront bien utiles pour tester des règles de phonétisation. En fait BDPHO est présenté sous deux formes compte tenu des contraintes d'exploitation (facilité, rapidité, taille mémoire). La première avec une carte HyperCard par entrée orthographique et l'autre pour laquelle un millier d'entrées sont regroupées par carte. Dans la première organisation, chaque unité de la base comporte six champs : la forme orthographique, le nombre total des occurrences, la (ou les) forme(s) phonétique(s) et la (ou les) cohorte(s) correspondante(s), le nombre des occurrences de celle(s)-ci, une indexation des 10 premières apparitions dans le corpus orthographique. Nous retrouvons des constantes bien mises en évidence dans l'histoire déjà longue des principales listes de fréquence (ZIPF, 1935), en particulier la régularité de la loi de distribution rang-occurrences et une observation bien connue : environ 50 mots permettent de couvrir 50% de l'ensemble des occurrences (52 mots pour le Trésor de la Langue Française avec ses 71 millions d'occurrences). Il est ainsi possible de relier les occurrences au rang par la relation :

nb d'occurrences = 10.75 – 1.25 ln(rang)

Il est intéressant de rapprocher les occurrences du corpus de celles de la base de données BRULEX [CONTENT, 90], dictionnaire orthographique qui comporte 35 746 entrées (sous-ensemble du Micro Robert). Il apparaît de façon manifeste que si les distributions sont comparables, elles sont systématiquement décalées : les tests du $\chi 2$ sont significatifs au seuil 0.001. Dans BDPHO il y a nettement plus de mots orthographiques, phonétiques

ayant un nombre de lettres compris entre 1 et 8, un nombre de sons entre 1 et 7, un nombre de syllabes de 1 à 2 que dans BRULEX, au-delà c'est l'inverse. Cette tendance s'affirmant dès le niveau orthographique, il est difficile de l'imputer à la façon de noter le schwa et elle semble bien révélatrice de la nature des deux corpus : dictionnaire orthographique pour BRULEX et corpus de français parlé pour BDPHO.

3. Un environnement convivial

Le choix de l'environnement a été guidé par un certain nombre de facteurs :

- un environnement de programmation souple et facilement évolutif qui ne soit pas réservé à des spécialistes de l'informatique,
- un minimum de contraintes commerciales pour la distribution (pas de licence, puisqu'il existe des versions *stand alone* d'HyperCard),
- la disposition en entrée et sortie de caractères API d'une bonne qualité typographique (polices vectorisées),
- un matériel évolué, standardisé et convivial.

Les données se présentent donc sous la forme de quatre piles. Une pile par corpus où chaque carte représente un ensemble de 150 lignes orthographiques numérotées avec pour chaque ligne sa transcription phonétique.

1 X / Président / de la Commission Nationale / d'aménagement / du Territoire /
x / prezi'dɑ̃ / də la kɔmisjɔ̃ nasjɔ'nal / d amenaʒ'mɑ̃ / dy teʀi'twaʀ /
2 Président / de la Compagnie / d'Aménagement / du barreau de
pʀezi'dɑ̃ / də la kɔ̃pa'ɲi / d amenaʒ'mɑ̃ / dy baʀo də

La pile BDPHO où chaque carte correspond à une entrée orthographique du lexique et où sont consignées dans les différents champs les informations qui la concernent (Fig.1).

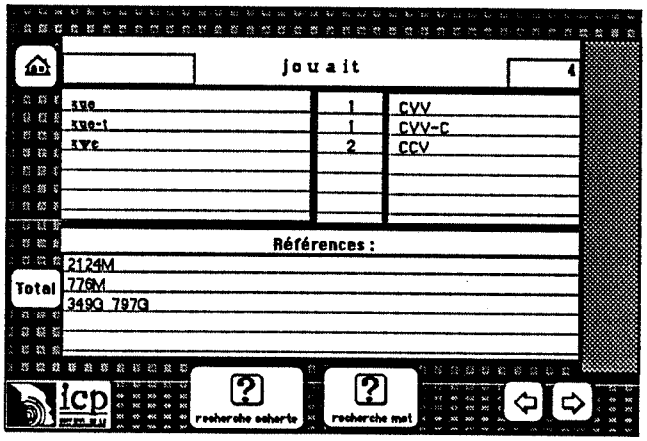


Fig. 1 : exemple de carte de la pile BDPHO
pour chaque forme rencontrée dans le corpus, nous avons tous les renseignements la concernant :
« 4 » : le nombre d'occurrences total
CVV, CVV-C, CCV : les cohortes
[ʒue], [ʒue-t], [ʒwe] : les variantes phonétiques
21254M, 776M, 349G, 797G : les références de la ligne où se trouvent les variantes phonétiques-corpus.

- L'environnement HyperCard et le langage HyperTalk (qui permet d'enchaîner des commandes d'HyperCard en les inscrivant dans le script attaché à chaque objet) rend aisé l'appel, à des logiciels extérieurs et la création de nouveaux champs informatifs.

Sur la carte courante de BDPHO les boutons recherche sur mot et recherche sur cohorte ont été programmés pour faciliter l'accès à une forme ou à une cohorte indiquée.

4. Le système PILAF

Ce système convivial permettant l'analyse du français (Procédures Interactives Appliquées au Français) a été mis au point par l'équipe TRILAN (Traitement Informatique de la Langue Naturelle).
Au niveau lexical, ce système propose des modules permettant l'analyse morphologique de textes et la génération de formes canoniques associées à une racine.

L'analyseur morphologique [COURTIN 92a] est un transducteur d'états finis de chaînes de caractères à chaînes de mots. Il calcule également les variables lexico-syntaxiques associées à ces mots. La réversibilité de ce transducteur permet la génération des formes associées à une racine en fonction d'un filtre défini par l'utilisateur. Ces modules utilisent une base de données composée de deux dictionnaires : un dictionnaire de bases, un dictionnaire de suffixes et désinences, et des données linguistiques comprenant une grammaire à validations et saturations définie par un ensemble de règles ainsi que des listes de modèles, de catégories grammaticales et de variables.

Deux modules permettent respectivement la mise au point interactive des dictionnaires et des paramètres linguistiques [COURTIN 91].

Le dictionnaire des bases lexicales, élaboré à partir du dictionnaire minimal de PILAF (bases correspondant aux modèles) et enrichi en utilisant la version électronique de lexique de M. GROSS [GROSS 93] contient environ 35 000 bases qui permettent la génération et la reconnaissance d'environ 250 000 formes.
Ce système PILAF a déjà permis [FRÉCHET 92a] le dépouillement d'un corpus oral (transcrit à l'écrit) dans le cadre d'une étude sur un langage opératif de communication homme-machine et la réalisation par le GETA (Groupe d'Études pour la Traduction Automatique) d'un lemmatiseur du Français [TOMASINO 90] intégré dans un outil d'indexation qui doit servir d'interface pour les bases documentaires. Et tout récemment, il a permis, dans le cadre du projet ORALL, la génération de toutes les formes dérivées des verbes issus du 60 000 ICP [BELRHALI 91] : près de 7 000 verbes ont été traités. Les données consistaient en une liste des verbes à l'infinitif suivi du code du LADL. Deux listes de résultats ont été fournies, l'une n'indiquant que les formes engendrées, l'autre chaque forme complétée des variables grammaticales correspondantes.

III. OBJECTIFS ET MISE EN ŒUVRE DE L'ÉTUDE

Les objectifs visent à obtenir une connaissance lexicographique plus approfondie des corpus. Pour cela, l'analyse ne doit pas seulement mettre en évidence les homographes (mots qui s'écrivent de la même façon et n'ont pas le même sens) mais encore les différencier

selon leurs variantes grammaticales. Pour chaque forme orthographique de la base BDPHO et pour chaque variante phonétique le but est d'indiquer quels sont les homographes présents dans les textes avec leur lemme et le nombre de fois où ils apparaissent. La mise en œuvre a été dictée par la difficulté d'obtenir des résultats fiables d'une analyse syntaxique des corpus donnés et d'avoir parfois même besoin de faire appel à la sémantique.

Les phrases sont parfois très longues et structurées fort différemment de celles des textes écrits. Elles présentent bon nombre de répétitions, d'hésitations et de mots inachevés :

La pile de lemmatisation devient ainsi le noyau central de l'étude.

Pour chaque carte de la pile, la lemmatisation de la forme orthographique est alors une application ponctuelle.

La création d'un bouton (Hylem) permet de lancer l'application de lemmatisation c'est-à-dire l'analyse morphologique de la forme et la lemmatisation de chaque homographe reconnu en utilisant des modules de PILAF ; les homographes lemmatisés apparaissent dans un champ texte de la carte. Pour préciser la présence ou non des homographes dans les corpus une liaison BDPHO-Corpus a été crée au moyen d'une Indexation qui permet d'afficher les lignes des corpus référencées.

Cet accès direct aux contextes de la forme orthographique permet de déterminer non seulement l'existence d'un homographe mais aussi de calculer le nombre de ses apparitions. Toutes les références seront indiquées et non plus seulement les 10 premières comme dans BDPHO.

Pour pallier à certains effets de bord dans le calcul précédent, une liaison permet d'afficher à tout moment, dans la fenêtre message, n'importe quelle ligne d'un des 3 corpus en indiquant simplement l'initiale du corpus et le numéro de la ligne souhaitée.

Une autre liaison dynamique BDPHO-PILAF qui permet d'apporter des modifications « en direct » en cas de non-conformité des résultats par rapport à ceux attendus, elle permet l'appel au logiciel PILAF et la mise à la disposition immédiate de tous ses modules.

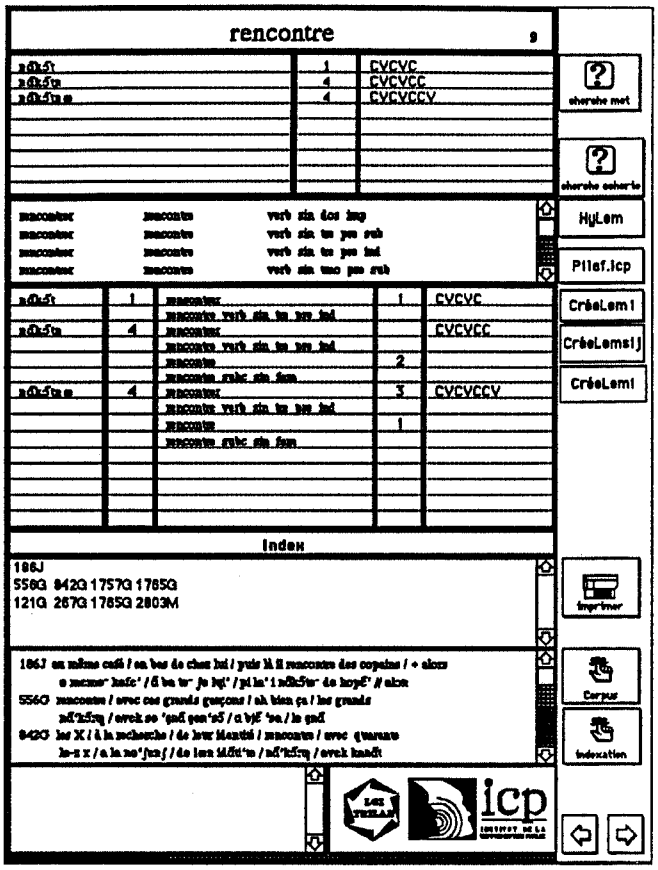
Dès qu'une carte de la pile BDPHO est affichée à l'écran, l'utilisateur a accès :

– à tout l'environnement HyperCard et peut donc en utiliser toutes ses commandes comme par exemple naviguer dans les piles pour faire des recherches ou pour apporter des corrections, créer de nouvelles piles résultats dont la fréquence est supérieure à n.

– à PILAF et à toutes ses commandes notamment à la commande de mise à jour du dictionnaire ou d'appel à l'aide.

Les boutons Créelemi ou Créelemsij ont été développés pour ne pas avoir à ressaisir les caractéristiques de tous les homographes, seul le numéro du lemme et sa fréquence seront indiqué à la demande. Les besoins en matériel ne sont pas considérables : le chargement d'HyperCard ne demande que 4 Mo de mémoire.

L'occupation du disque peut varier entre 4 et 12 Mo. Le système HyperCard 2.1 et sa pile de base n'occupent respectivement que 702 ko et 50 ko.



Si le dictionnaire de PILAF a besoin de 1149 ko, il suffit de 48 ko pour les grammaires. La place occupée par l'application PILAF n'est que de 140 ko et celle occupée par l'application Hylem seulement 35 ko. Les trois corpus requièrent 1503 ko c'est-à-dire, respectivement 498 ko pour le corpus G, 99 ko pou J et 906 ko pour M.

La place occupée par la pile BDPHO originelle est de 650 ko, celle occupée par la nouvelle pile après indexation et lemmatisation d'environ 7 400 formes orthographiques de fréquence inférieure ou égale à 20 est de l'ordre de 8 Mo. Il n'est nullement besoin d'avoir la disponibilité de cette place, car l'indexation peut être faite à la demande et en fonction de l'étude menée, des mini-piles ne concernant que les formes à étudier peuvent être créées.

IV. PREMIERS RÉSULTATS

Parmi les 7597 formes de la pile initiale BDPHO, 850 avaient une fréquence supérieure à 10, 2951 une fréquence comprise entre 1 et 10, et 3821 formes avaient une fréquence égale à 1.

La stratégie pour traiter les formes ayant une fréquence supérieure à 50 n'est pas encore tout à fait arrêtée. Ces formes correspondent, à quelques rares exceptions, à des mots-outils et un premier regard sur les résultats fait apparaître que des modifications sensibles risquent d'intervenir au niveau de leur fréquence.

Nous avons recensé et regroupé en 6 principales catégories les différentes questions soulevées au cours de ce travail :

• Les lettres accentuées, qui transcrites en majuscules perdent leur accentuation. Par exemple A est-il être une occurrence de "à" ou de "a" ? S'il est considéré comme une occurrence de "a", la lemmatisation de cette forme doit alors faire apparaître les deux homographes,

à, a préposition, et a, a auxiliaire avoir

ou alors être occurrence de la forme "A" qui elle seule présentera les deux homographes. S'il y a création de la forme "A", le traitement de la forme "à" devient aisé malgré sa fréquence élevée puisqu'il n'y a plus d'homographe, mais alors il faut que les formes « Etat », « Ile »... soient, elles aussi, reconnues et non rattachées aux formes accentuées.
Pour ce problème et grâce à l'évolution des polices de caractères nous avons décidé d'accentuer les lettres majuscules dans les corpus et de les dénombrer avec l'occurrence transcrite en minuscules.

• Les locutions qui, non reconnues en tant que forme orthographique sont recensées à chacun de leurs constituants. Quelle catégorie faut-il alors assigner à ces derniers comme par exemple à « prou », « extenso », « priori » dans les locutions adverbiales « peu ou prou », « in extenso », « a priori »?
Comme l'espace n'est pas considéré comme un séparateur par le système PILAF, nous avons décidé de regrouper en une seule forme orthographique, outre les locutions adverbiales, les locutions prépositionnelles et les locutions conjonctives. L'environnement HyperCard permet de créer facilement les cartes correspondantes de ces locutions et de modifier ou de supprimer les cartes de leurs constituants.

• Les mots composés communs ou propres tels « fume-cigarette », « bateau-mouche », « avant-garde », « états-majors ».

Actuellement il existe une entrée « majors » comment lemmatiser cette forme ? En effet « major » en français est un adjectif invariable, de même pour « Olonnes » ou « Gaulle » dans les « Sables d'Olonnes » ou « De Gaulle ». Il nous a semblé normal d'harmoniser la solution avec celle prise pour les locutions et de considérer les mots-composés comme une seule entité.

• Les préfixes qui, en général sont moins employés à l'oral qu'à l'écrit semblent fort peu nombreux et fort peu utilisés dans ces corpus. Suite logique aux solutions précédentes, nous avons regroupé sur une même entrée la forme préfixée.

• Les sigles et les abréviations
Pour ces dernières, l'obtention de la forme complète pour le lemme paraît possible après la création d'un modèle spécifique et quelques modifications de Pilem ce qui donnerait pour les abréviations les résultats suivants :

agrégation	agreg	subc	fem sin
publicité	pub	subc	fem sin

Le même principe pourrait être retenu pour les sigles mais alors la longueur de la chaîne représentant le lemme ne pourra dépasser 25 caractères, taille maximum retenue dans PILAF pour la longueur d'un mot, longueur très facilement atteinte.
HLM Habitation à loyer modéré
Pour les problèmes rencontrés, en ce qui concerne les catégories, nous avons pris comme référence Le Trésor de La Langue Française, et Le Robert de La Langue Française. Nous avons ainsi créé deux nouvelles catégories lexicales "Abrv" et "Sigl" (abréviations, et sigles)
En l'état actuel, l'étude a permis :

• La réalisation par I. KOWARSKI du lemmatiseur Hylem avec l'intégration d'un module au système PILAF.

• L'ajout de mots absents du dictionnaire de PILAF : « distale », « orientalisante », « prosodème », « racketter », « nullard »

• La détection de quelques erreurs dans la pile : des cartes répertoriées deux fois avec des références différentes : « économie », « école »

• La détection d'erreurs lors de la transcription orthographique des corpus (pour la plupart corrigées dans l'édition du « De À à Zut » mais non reportées dans les corpus) et leur correction immédiate.
Ces erreurs qui sont en majorité

– des redoublements de consonne : appellait appele, apperçoive, aprennent, bouillonement, incessament

– des erreurs d'accentuation : acheterai, apparaîssait, connait, dûe, élévé, lancait

– des fautes de frappes : actuellemt concacrée, dangeureuse, dixaine, incabable, locomitive, maintanant

– des fautes d'orthographes : asseoit fatiguante piedmontais.

– quelques erreurs d'accord : on donnent, des feu de bois, leur bureaux, leur parents, on avait parler.

• La détection de quelques erreurs dans le « De A à Zut » : fénéants, marais

• La détection "par hasard" de fautes plus difficilement détectables lors d'une lecture en ligne des corpus.

2235M	ça c'est l'ennuie
1482G	puisque la marais a monté brusquement
2074M	Cela n'existe nul par ailleurs

V. CONCLUSION

Ce travail qui fait suite à l'élaboration du dictionnaire du français parlé « De À à Zut » avec l'obtention des formes lemmatisées, a permis une correction fine des corpus. Au passage, il a permis de confirmer l'adaptabilité informatique et linguistique du système PILAF et la robustesse de son dictionnaire de bases mais a montré sa faiblesse sur le plan syntaxique pour ce type de corpus.

La version 2.0 de la pile BDPHO et des corpus corrigés, en fournissant un ensemble cohérent et convivial des données deviennent un instrument agréable à manipuler et qui semble utile aussi bien au linguiste qu'à l'informaticien et au phonéticien.

La simultanéité des données, des résultats et des contextes permet une approche nouvelle des méthodes de travail.

La poursuite des travaux peut être orientée vers :

• L'étude et le recensement des formes orthographiques qui donnent les mêmes homographies.

rencontre (verbe) / rencontre (substantif)

Cette classe d'ambiguïté homographique et catégorielle semble constituer un ensemble très important dans BDPHO.

• La détermination d'un analyseur syntaxique plus adapté à de tels textes. Pour cela, il peut être intéressant de pouvoir appeler ponctuellement divers analyseurs sur les différents contextes d'une même forme et d'en comparer aisément les résultats, d'étudier les effets de distance et peut-être d'en dégager un ensemble de structures syntaxiques minimales en fonction des catégories lexicales pour mieux prendre en compte les répétitions, les mots inachevés, les interruptions...

• L'étude des ambiguïtés apportées par la retranscription orthographique de l'oral :

nous portions (verbe) des portions (nom)

président (verbe) et président (nom)

Une étude plus spécifique du français parlé est déjà engagée. Après avoir constaté, l'émergence de formes caractéristiques de l'oral :

ça (1069), euh (977), y (843), alors (552), très (517), oui (499), enfin (491) parce que (480) moi (345), quand (289), puis (263), ah (217), dont la fréquence est nettement supérieure à la fréquence de ces mêmes formes écrites, certains de ces mots semblent pouvoir jouer un rôle de marqueurs syntaxiques :

• « euh », « bon », « ben » peuvent être considérés comme des pauses sonores qui servent à maintenir le contact entre le locuteur et son destinataire, lorsque le locuteur veut garder la parole tout en réfléchissant à ce qu'il va dire. Leur présence peut :

• correspondre à une ponctuation de l'écrit et introduire un nouveau groupe syntaxique :

30 M enfin ils m'ont laissé du temps pour finir mon Doctorat de Troisième Cycle euh / je fais un peu mes recherches pour la Thèse d'Etat enfin à côté de mon travail qui est de la phonétique euh / enfin de la phonétique pédagogique /

• faire recourir à la reprise d'une structure équivalente

48 M d'une part je fais une euh / j'ai fait une recherche sur les / les consonnes finales les occlusives en finale

50 M l'aspect de / ce que nous appelons euh / ce que l'on appelle / de façon générale l'aspiration /

• n'avoir aucune influence sur la structure en cours

23 M Monsieur X / parce que X est maintenant euh / à Toulouse enfin il est /.

• Des expressions comme "si vous voulez", "remarquez", "vous comprenez", "nous voyons" semblent, elles aussi, jouer un rôle de pause rythmique.

854 G de sous-titres / si vous voulez / qui éclairent / le texte // Je voudrais

1630 M sciences plus appliquées / quoi/ remarquez euh / le but qu'on recherche tous / en

3408 M cet étonnement / cet amusement / vous comprenez / de / de l'étonnement / implique /

La création de nouvelles classes grammaticales devraient permettre une description lexicale et syntaxique de ces corpus plus appropriée.

BIBLIOGRAPHIE

[BELRHALI 91] : Rabia BELRHALI, Laure LIBERT, *Règles de phonétisation d'un lexique de 55 000 mots.* Mémoire de DEA, Université Stendhal, 1991.

[BOË 92a] : Louis-Jean BOË, Jean-Pierre TUBACH, *BDPHO : une base de données lexicale orthographique-phonétique du français parlé,* Séminaire Lexique du GRECO-PRC Communication Homme-Machine, Toulouse, 111-119, 1992.

[BOË 90b] : Louis-Jean BOË, Jean-Pierre TUBACH, *«De À à Zut» Dictionnaire phonétique du français parlé,* Editions Ellug, Grenoble, 1992.

[CATACH 84] : Nina CATACH, *Les listes orthographiques de base (LOB),* Nathan, Paris, 1984.

[CONTENT 90] : Alain CONTENT, Philippe MOUSTY, Monique RADEAU, *BRULEX une base de dennées lexicales informatisée pour le français écrit et parlé.* L'Année Psychologique, 551-556, 1990.

[COURTIN 91] : Jacques COURTIN, Danièle DUJARDIN, *Paramètres linguistiques du français dans le système PILAF.* Rapport Technique RT 67, Laboratoire de Génie Informatique, Grenoble, 1991.

[COURTIN 92a] : Jacques COURTIN, Danièle DUJARDIN, Damien GENTHIAL, Irène KOWARSKI, *Outils lexicaux de l'équipe TRILAN : bilans et perspectives.* Séminaire Lexique du GRECO-PRC Communication Homme-Machine, Toulouse, 93-109, 1992.

[COURTIN 92b] : Jacques COURTIN, Danièle DUJARDIN, Irène KOWARSKI, PILAF : *Software Tools for Lexicography and Text Research.* COMPLEX'92: 2nd Conference on Computational Lexicography and Text Research, Budapest, Hungary, 93-109, 1992.

[FRÉCHET 92a] : Anne-Lise FRÉCHET, Danièle DUJARDIN, Marie-Annick MOREL, Jean CAELEN, *Analyse lexicale pour le français oral avec le logiciel PILAF.* Séminaire Lexique du GRECO-PRC Communication Homme-Machine, Toulouse, 239-251, 1992.

[FRÉCHET 92b] : Anne-Lise FRÉCHET, Marie-Annick MOREL, Danièle DUJARDIN, Jean CAELEN, *Analyse lexicale d'un langage opératif,* Bulletin de la Communication Parlée, 2, 167-182, 1992.

[GROSS 93] : Maurice GROSS, Max SILBERZTEIN, *Outils de traitement linguistique, applications à l'analyse documentaire.* Actes de la troisième école d'été, Lannion, 1993.

[ENGWALL 84] : Gunnel ENGWALL, *Vocabulaire du roman français (1962-1968). Dictionnaire des fréquences,* Almqvist & Wiksel International, Stockholm, Suède, 1984.

[TOMASINO 90] : Isabelle TOMASINO, *ODILE, un outil d'intégration de dictionnaires et de lemmatiseurs.* Thèse CNAM, Grenoble, 1990.

[VAN EIBERGEN 85] : Joëlle VAN EIBERGEN, *Corpus d'un français vernaculaire à caractère spontané et impératif,* Bulletin de l'Institut de Phonétique de Grenoble, 15, 35-74, 1985.

[ZIPF 35] : Georges K. ZIPF, *The Psycho-Biology of language,* Hough, Houston, 1935.

SAUVEGARDE LEXICALE ET PHONÉTIQUE DE DIALECTES ORIGINE DES TOPONYMES

André FRANCÈS & Rabia BELRHALI

Centre de dialectologie de l'Université Stendhal

Résumé :

Au cours de nombreuses années d'enquêtes auprès des patoisants de certaines communes, nous avons décidé de nous intéresser plus particulièrement à la recherche de la signification des toponymes à l'aide du vocabulaire qui nous a été fourni. La saisie des données s'est fait dans un environnement HyperCard (Macintosh) afin de permettre leur exploitation. L'objectif de cette étude est de sauvegarder des éléments oraux des "langues du pays", mais aussi de rapprocher des noms communs de noms propres que sont les toponymes afin d'en trouver la signification.

Mots clés : Toponymes, enquêtes, patois ("langues de pays").

INTRODUCTION

Les noms de lieux, toponymes, posent un réel problème. D'où viennent en effet ces noms de lieux qui nous sont si familiers et dont on ignore la signification ? Pour répondre à cette question, nous nous sommes fixés deux objectifs :

• Sauver – en les notant par écrit – les éléments jusqu'ici purement oraux du vocabulaire courant des « langues du pays » ou « patois ». Sont donc enregistrées les réponses des derniers patoisants qui habitent des localités relativement à l'écart dans une zone de 100 km environ autour de Grenoble ainsi que dans la zone d'influence de Briançon (Hautes Alpes).

• Tenter de rapprocher certains de ces noms communs, des noms propres que sont les noms de lieux de ces mêmes localités, pour essayer de retrouver leur signification – aujourd'hui perdue la plupart du temps.

Le premier axe de recherche fait appel à une gestion informatisée ; par son caractère propre cette technique apporte une grande facilité dans la saisie et dans l'exploitation d'une masse de données qu'il serait difficile et fastidieux de traiter manuellement. Il s'agit en définitive du traitement des données brutes – les termes de sens commun –, chacune de ces données ne faisant l'objet que de commentaires limités.

En revanche le deuxième axe de recherche ne fait appel à cette gestion informatisée qu'accessoirement par un simple rapprochement suggéré entre toponyme et nom commun enregistré.

Nous nous proposons dans une première partie de présenter les modalités des enquêtes ; l'exploitation du corpus fera l'objet d'une deuxième partie. Enfin à travers des exemples nous expliquerons quelques toponymes.

I. MODALITÉ DES ENQUÊTES

Ces enquêtes menées par A. Francès ont débuté il y a une quinzaine d'années. Elles se poursuivent toujours mais à un rythme ralenti pour deux raisons :
– les témoins patoisants, nécessairement âgés, sont de plus en plus rares,
– la masse de documents devient suffisante pour une exploitation sélective.

Une étude de la configuration géographique (géomorphisme) de chaque lieu dit (rocher, rivière, végétation...) permettra à l'enquêteur de réaliser une collecte de données. La prononciation du toponyme par les patoisants est demandée afin de permettre une comparaison entre la réalisation phonique correspondant au toponyme étudié. Cette comparaison permettra d'expliquer des toponymes dont le sens n'est pas clair au départ.

Les mots patois sont notés en utilisant le système Gilliéron sur des feuilles doubles numérotées, datées avec le nom du témoin. Ainsi un dossier est ouvert par commune constituant un point d'enquête.

Tous les points enquêtés ne l'ont pas été à un degré similaire :

• Monêtier les Bains (vallée de la Guisane) à mi-chemin entre le Col du Lautaret et Briançon, a fait l'objet d'une investigation plus dense que les autres. A. Francès y disposait de facilités qui lui ont permis d'engranger, au cours d'enquêtes étalées dans le temps, 3 000 à 4 000 mots ou expressions de sens commun dont la moitié est saisie informatiquement à l'heure actuelle.

Viennent ensuite des communes où une moisson substantielle a pu être engrangée (présence de témoins cultivés, disponibles et curieux d'esprit) :

• Chichilianne (Trièves) : 700 à 800 mots ou expressions,

• Les Avanchers (Savoie), et Cervières (Hautes Alpes) : 700 à 800 mots ou expressions.

• Ceillac (Queyras), la Salle les Alpes (Hautes Alpes), Saint Christophe en Oisans (Isère), 500 mots environ chacune. Dans ces trois dernières communes la qualité supplée à la quantité : on a pu y recenser de vieux mots

sonnant "l'ancien" patois devant l'inévitable francisation rampante.

Pour les points où A. Francès a eu des facilités d'enquêtes, il a été demandé non seulement le vocabulaire « toponymique », mais aussi le vocabulaire courant.

Dans les communes où les possibilités de récolte étaient plus faibles, une priorité a été réservée au vocabulaire dit toponymique : ce vocabulaire couvre les thèmes que traduisent généralement les noms de lieux. Ces thèmes sont généralement :
– le degré de fertilité de la terre
– la topographie générale
– la nature des sols
– le relief...

II. MÉTHODOLOGIE ET MATÉRIEL UTILISÉ

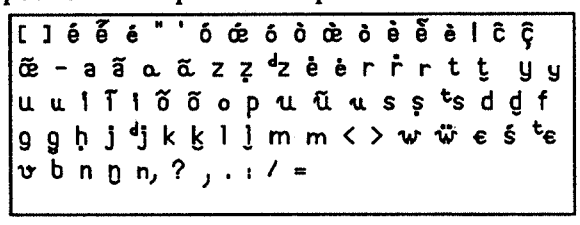
Immanquablement allait se poser la question de l'exploitation de cette masse de données manuscrites donc difficilement transmissibles.

Le système des fiches cartonnées, manuscrites, rangées dans des contenants style "anciennes boîtes à chaussures" n'est plus au goût du jour; il a bien été employé autrefois par d'infatigables et passionnés chercheurs auxquels il faut rendre hommage pour leur ténacité : le fichier, bien connu des provençalistes, de l'Abbé Moutier sur les parlers de la Drôme Provençale en est l'exemple type.

À la bibliothèque municipale de Valence, nous avons pu admirer la somme de labeur que représente la trentaine de boîtes de chaussures, bourrées à refus de petites fiches manuscrites !
Afin de gérer, stocker nos notes manuscrites, nous avons fait appel à un système informatique.

Notre choix s'est porté sur un Macintosh (SE 1 / 40). Le choix de l'environnement HyperCard nous semble répondre à nos besoins, d'autant plus que celui-ci n'est pas réservé uniquement à des spécialistes de l'informatique. La gestion de nos données sous forme de piles nous paraît totalement satisfaisante.

Nous avons dans un premier temps créé une police de caractères type Gilliéron en conformité avec nos notes manuscrites, car aucune police de caractères Gilliéron n'était disponible dans le commerce. Cette police de caractères, dessinée avec le logiciel Fontastic Plus et installé dans le système avec l'utilitaire Font DA Mover, existe en deux corps de caractères possibles : corps 12 et corps 14.



Police de caractères Gilliéron : corps 12

Le système de signes phonétiques Gilliéron a été préféré au système international : outre ses mérites propres, c'est le système utilisé
• par l'atlas linguistique de France (ALF) [1]
• par certains atlas linguistiques régionaux :
– l'atlas du Jura et des Alpes du Nord (ALJA) [2]
– l'atlas de Provence (ALP) [3]
(les points d'enquête sont situés dans les deux zones couvertes par l'ALJA et l'ALP).
• par des glossaires spécialisés tels que le glossaire des patois francoprovençaux d'Antonin Duraffour en particulier [4].

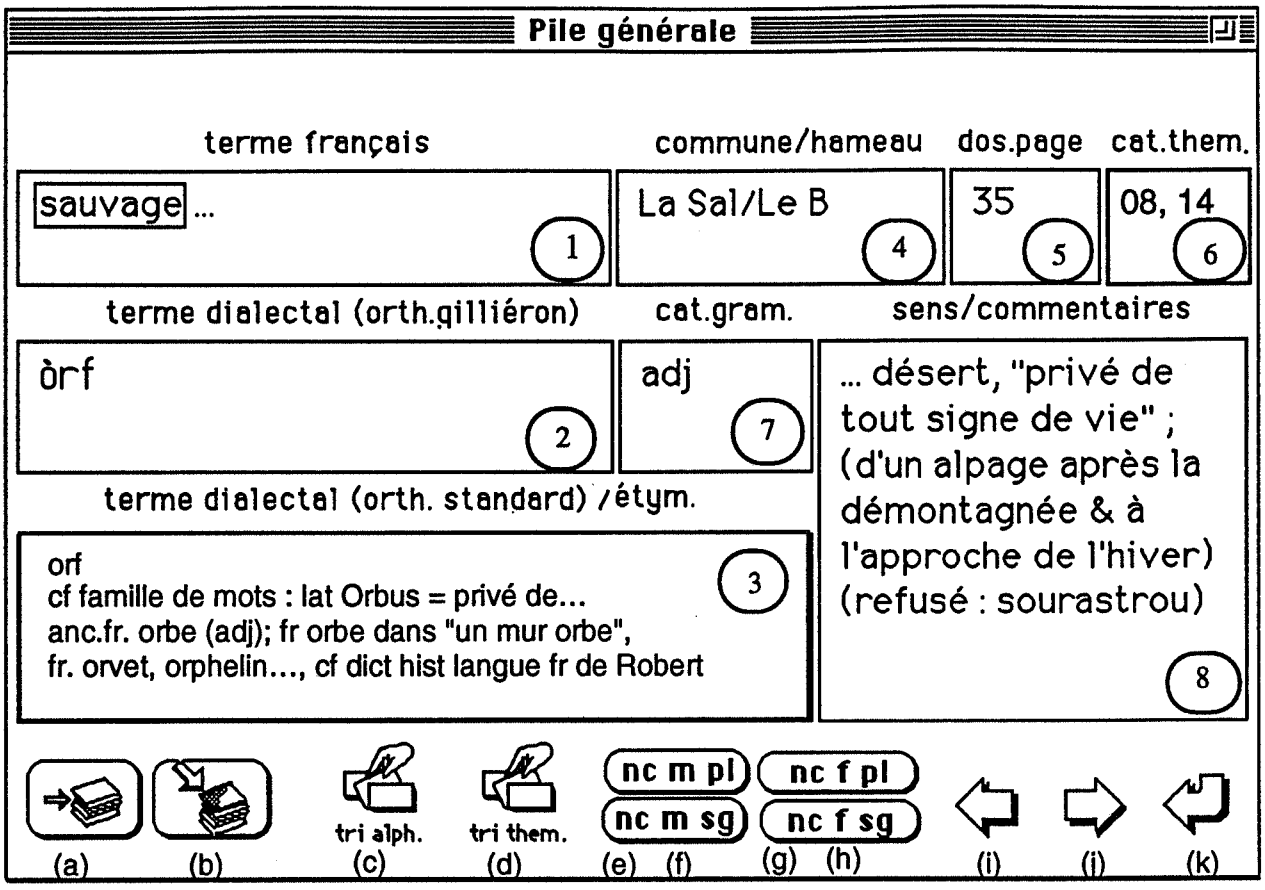
Les enquêtes sont menées suivant un "fil conducteur" ou plutôt des "fils conducteurs" qui peuvent varier d'un point à un autre. Ainsi pour Monêtier les Bains (Vallée de la Guisane dans les Hautes Alpes) a été utilisée la thèse de l'abbé Jean Delaigue [5] relative au patois d'Araules près d'Yssingeaux dans le Massif Central : le patois d'Araules est proche de celui de Monêtier les Bains ; ailleurs comme à Cervières (vallée de la Cerveyrette dans les Hautes Alpes), aux Avanchers (Savoie) le questionnaire de G. Tuaillon a été utilisé. Ce questionnaire a été mis au point au Centre de Dialectologie de l'Université Stendhal de Grenoble précisément pour les enquêtes dialectologiques en pays alpin.
Il faut mentionner aussi les vocabulaires plus spécialement centrés sur les thèmes toponymiques qui accompagnent divers ouvrages :
– le dictionnaire des patois du Queyras de Chabrand et de Rochas d'Aiglun [6]
– les ouvrages récents d'Hubert Bessat et Claudette Germi [7], [8]

Dans un premier temps, nous avons créé un cahier des charges pour mettre au point une carte type qui répondrait aux besoins pour ce genre de travail.

L'archivage des données se fait en deux étapes.
• La première étape consiste à accumuler, toutes communes confondues, les documents traduisant les informations manuscrites dans une pile générale.
• La deuxième étape consiste à créer, à partir de la pile générale, des mini-piles par commune.

La pile générale est constituée d'une première carte qui contient 30 champs. Chaque champ annonce la commune d'enquête. À partir de cette carte, en cliquant sur le champ de la commune considérée, on crée une mini-pile où seront transférées automatiquement toutes les cartes concernant la commune.

La deuxième carte de la pile générale est la carte type qui va servir de fond de carte. Elle apparaîtra à l'écran autant de fois qu'il y aura un mot ou une expression à répertorier.



exemple de carte

Dans cet exemple de carte, il s'agit de la commune de la Salle les Alpes / hameau du Bez, sigle La Sal/Le B.

Cette carte illustre le cas où nous recherchions un mot de vocabulaire courant sans recherche toponymique ultérieure.

Nous avons demandé l'équivalent patois du mot « sauvage » dans l'expression « endroit sauvage, désert, privé de vie », ce qui se produit après la "démontagnée" d'un alpage.

La réponse obtenue fut l'adjectif « orf » à rapprocher d'une famille de mots comprenant :

– le latin *orbus* dans le sens : privé de ...

– l'adjectif d'ancien français « orbe » dans le sens : aveugle, privé de vue.

– les mots de français contemporain :

« orbe » dans « mur orbe », c'est-à-dire « mur aveugle sans ouverture ».

Nous avons noté également les dérivés « orvet, orphelin » ainsi que les références bibliographiques.

• Dans le champ (1) : on inscrit le terme français, mot ou expression, demandé au témoin.

• Dans le champ (2) : le terme dialectal est transcrit en phonétique Gilliéron

• Dans le champ (3) : le terme dialectal est repris mais cette fois en orthographe standard, suivi éventuellement de considération étymologiques.

La réponse du témoin est transcrite deux fois pour deux raisons :

1° – l'écriture en caractère standard est utile pour le classement des fiches : le bouton tri alphabétique permet le classement alphabétique soit à partir du terme français soit à partir du terme dialectal standard mais pas à partir du terme dialectal Gilliéron.

Mais inconvénient de la graphie standard : on ne peut matérialiser la voyelle accentuée. En revanche la notation de l'accentuation est possible avec le terme Gilliéron : est accentuée la dernière voyelle écrite en caractère normal.

2° – la restitution des sons est plus fidèle avec une orthographe phonétique.

• Le champ (4) : indique le nom de la commune ou de hameau où a été noté le mot, l'expression.

Chaque enregistrement est fait dans la pile générale. On travaille en pratique par séquence relative à une commune, on n'est plus obligé d'enregistrer la totalité des témoignages reçus dans une commune avant de prendre la suivante, on peut changer de communes en cours de travail.

• Dans le champ (5) : est indiquée la page où se trouve le témoignage dans une note manuscrite relative à la commune considérée, indication précieuse pour des recherches comparatives entre communes.

• Le champ (6) : contient la catégorie thématique du mot ou de l'expression

• Dans le champ (7) : nous précisons la catégorie grammaticale du terme dialectal : nom, adjectif, verbe...

• Et enfin dans le dernier champ (8) : nous précisons la définition du terme demandé, nous ajoutons quelques commentaires au besoin, des références d'emploi.
• Le bouton noté (a) dans la figure 1, nous permet de passer de la pile générale à la mini pile correspondant à la commune sur laquelle le travail s'effectue.
• Le bouton (b) nous permet :
– d'« envoyer » une carte nouvellement créée dans la mini pile de la commune considérée.
– de remplacer, après modifications, une carte existante dans la mini pile à partir de la pile générale.
Les modifications que l'on veut apporter à une carte doivent toujours s'effectuer à partir de la pile générale. Ceci afin d'obtenir une pile générale qui soit la somme des mini piles pour permettre d'éventuelles recherches sur toutes les communes.
• Les boutons (c) et (d) servent à enclencher le tri alphabétique ou thématique des cartes de la pile générale.
• En cliquant sur l'un des boutons (e), (f), (g), (h), on inscrit automatiquement l'expression correspondant dans le champ catégorie grammaticale.
• Les boutons (i) et (j) servent à tourner les cartes de la pile générale une à une dans un sens ou dans l'autre.
• Et enfin, le bouton (k) nous permet de revenir à la première carte de la pile générale.

À l'alinéa "précisions sur les axes de recherches" il a été fait allusion à de simples et sporadiques rapprochements entre certains termes communs et des toponymes de la commune étudiée.
Ces rapprochements sont présentés et commentés ci-dessous dans le cadre des cartes types dont l'empilement constitue, d'une part la pile générale relative à toutes les communes confondues, d'autre part les piles particulières à chaque commune d'enquête.

III. APPROCHE DE LA SIGNIFICATION DES TOPONYMES À TRAVERS QUELQUES EXEMPLES

EXEMPLE 1

À Monêtier, il y a la montagne de l'Area (forme française mais « Aryas » en patois). Que signifient ces mots Area, Aryas ?

Il se trouve que l'adjectif commun « aride » se dit « arya » en patois : cette terre, elle est aride :

[arya / ikèla tèra, i l i arya][1]

La parenté – en patois – entre l'appellatif et le toponyme est appuyée par une caractéristique nette des lieux : des landes sèches, dures où ne poussent qu'arbustes xérophiles, occupent les flancs de la montagne.

Il faut donc admettre que la parenté phonétique correspond à la parenté sémantique. La signification commune du toponyme jusqu'ici perdue dans l'esprit des autochtones est rétablie : le toponyme Area (en français) ou Aryas (en patois) signifie : la montagne aux flancs arides.

EXEMPLE 2

Exemple où une règle phonétique bien connue conduit à émettre une hypothèse toponymique dans un champ d'enquête étendue :
– à Monêtier les Bains « écorce résineux » se dit en patois [ɾuᵗɛa][1] qui vient du latin *rusca*. Tenons-nous l'explication du toponyme Ruchère en Chartreuse (Col de la —) en français, mais « rustire » en patois de Chartreuse ? Un C latin suivi d'un A, en position forte, aboutit au son [ʃ][2] - en français, [tʃ] - à Monêtier, [st] - en Chartreuse. Il nous restait à vérifier si le col n'est pas un lieu d'un chantier d'exploitation de la forêt.

EXEMPLE 3

À Monêtier deux lieux dits intriguaient les autochtones – Mandette, Mandoline – qui cherchaient en vain à expliquer l'allusion à un instrument de musique en ce lieu.

En patois d'Embrun, il existe un terme de sens commun « muande » qui désigne un chalet d'alpage de séjour temporaire pour les alpagistes accompagnant les troupeaux dans leur montée aux alpages supérieurs (du latin *mutanda* = (lieux de séjour) devant être changés).

Le toponyme Muande – de Monêtier les Bains – est assimilable au « muande » d'Embrun (diphtongaison ou non de la voyelle accentuée).

La situation des deux toponymes est compatible avec cette assimilation.

À noter que moyennant l'insertion d'un r- supplétif et non étymologique on peut interpréter de même le toponyme de la Chartreuse : l'Emeindra. Emeindra qu'il faut comprendre « le meindra », formes féminines au pluriel, qui signifient « les chalets d'alpage temporaire ».

EXEMPLE 4

Dans la commune de Monêtier les Bains, nous cherchions à expliquer le toponyme « Fossa » (prononcé [fusa] en patois). Ce lieu est situé entre Puy Saint Pierre et Puy Saint André dans le Briançonnais. Une enquête a permis de rapprocher le nom commun patois « foussa » qui signifie « fossé » et le toponyme recherché. La parenté phonétique entre l'appellatif en patois et le toponyme, conjuguée à une description topographique conforme n'a posé aucune difficulté d'interprétation.

EXEMPLE 5

Nous avons demandé l'équivalent patois de « orge ». Nous avons obtenu [ɯ̈ɛrᵈjɛ] en patois. Nous remarquons que certains lieux dit : les Orgières (commune de Saint Martin de Queyrières (Briançonnais) et du Bourget (Maurienne)), ne relèvent pas du thème de l'orge car leur situation en haute altitude est incompatible avec la présence de l'orge. On expliquera ces toponymes par le mot latin *horrearias* qui signifie « endroit des fenils »

[1] Caractères Gilliéron

[2] API

Nous rappelons, au passage, une règle de phonétique historique : R latin suivi d'un yod aboutit au son [ʒ]; ce qui permet l'interprétation correcte de plusieurs toponymes : Orgières (qui signifie les greniers d'altitude ou fenils).

EXEMPLE 6

Le terme patois « apra » à Monêtier signifie « âpre, difficile à cultiver » (en parlant d'un terrain). On peut rapprocher ce nom commun du toponyme « La Pra » car cette dernière forme résulte visiblement d'une fausse coupure. Le terme « pra » en patois qui veut dire « pré » étant du genre masculin est écarté. « Le pré » aurait comme équivalent « le pra » et non « la pra ». Il faut donc rétablir la forme correcte du toponyme à savoir « l'apra » c'est-à-dire les terrains difficiles à cultiver. « L'apra » signifie « les lieux âpres ».

EXEMPLE 7

À La Salle / Le Bez « maya » est un verbe patois signifiant « brouter » (du latin *malleare* qui veut dire actionner les *malleus* c'est-à-dire les maillets ou les molaires). Cette attestation nous donne la signification de nombreux toponymes : « Les Mayes », jusqu'ici énigmatiques, car les autres étymons latins aboutissant à May – à savoir *maccula, medalia, magide, maiu* – sous-tendaient des thèmes incompatibles avec la description topographique des toponymes considérés.

Les Mayes de Saint Christophe en Oisans, de l'Alpe de Venosc, du Val Jouffrey..., sont les seuls endroits herbeux et facile d'accès dans un paysage fait par ailleurs de pierres et de roches.

CONCLUSION

Ce travail est en cours d'exécution. Nous pouvons donc espérer disposer dans quelques temps d'un vocabulaire substantiel centré sur des thèmes toponymiques dans la majorité des cas, et sur des thèmes plus élargis dans le cas de Monêtier les Bains. Au-delà de l'aspect documentaire, nous espérons illustrer par de nombreux exemples une méthode dans l'approche de la signification des noms de lieux que sont les toponymes.

BIBLIOGRAPHIE

[1] GILLIÉRON J. & EDMONT E. ; *Atlas linguistique de la France,* Honoré Champion Éditeur, Paris, 1902.

[2] MARTIN J.B & TUAILLON G. ; *Atlas linguistqiue et éthnographique du Jura et des Alpes du Nord.* Édition du CNRS, Paris, 1971.

[3] BOUVIER J.C. & MARTEL Cl. ; *Atlas linguistique et éthnographique de la Provence,* Paris, 1975.

[4] BESSAT J.C & GERMI Cl. ; *Les mots de la montagne autour du Mont Blanc,* ELLUG, Université Stendhal, Grenoble, 1991.

[5] ABBÉ DELAIGUE J. ; *Glossaire méthodique du patois Aranles Haute-Loire ;* exemplaire dactylographié à la bibliothèque des facultés catholiques de Lyon, 1965.

[6] CHABRAND J-A. & DE ROCHAS D'AIGLUN A. ; *Patois des Alpes Cottiennes et en particulier du Queyras,* Édition Laffite Reprint, Marseille, 1973.

[7] BESSAT J.C & GERMI Cl. ; *Lieux en mémoire, de l'Alpe,* ELLUG, Université Stendhal, Grenoble, 1993.

[8] DURAFFOUR A. ; *Glossaire des patois francoprovençaux,* Édition du CNRS, Paris, 1969.

TRANSCRIPTION MULTINIVEAU D'UN CORPUS DE PAROLE

Alix Mailland, Martine de Calmès, Guy Pérennou

IRIT, URA CNRS 1399, Groupe IHM-PT
Université Paul Sabatier, 118 route de Narbonne
31062 Toulouse CEDEX
Tél. : 61 55 65 78, Fax : 61 55 62 86, e-mail : perennou@irit.irit.fr

RESUME

Nous nous proposons de décrire la version actuelle de la composante phonologique de BDLEX ainsi que les résultats obtenus sur une partie du corpus BREF80 développé au LIMSI dans le cadre du projet BREF.

L'approche présentée ici, basée sur les groupes à prononciations multiples (gpm) et les groupes phonologiques contextuels (gpc), permet de générer une transcription phonotypique unique à partir d'une transcription orthographique. La transcription phonotypique rend compte des prononciations multiples d'un énoncé.

Ce travail est développé dans le cadre du pôle parole du GDR-PRC CHM. Il s'effectue en liaison avec le projet BREF mis en œuvre par le laboratoire LIMSI.

MOTS-CLES

Traitement automatique de la parole, composante phonologique, apprentissage.

1. INTRODUCTION

Cette communication décrit le modèle phonologique développé à l'IRIT équipe IHM-PT. L'intérêt de notre composante phonologique est d'effectuer la transformation d'une forme orthographique en une forme phonotypique unique qui rend compte de toutes les prononciations possibles de l'énoncé.

Nous présentons tout d'abord le modèle phonologique qui utilise les unités phonologiques gpm et gpc ainsi que l'apprentissage effectué sur les règles phonétiques. Puis nous décrivons les observations faites sur un sous-corpus de BREF80 développé au LIMSI dans le cadre du projet BREF.

2. LE MODELE PHONOLOGIQUE

Dans la plupart des applications en traitement automatique de la parole et des textes il est préférable de confier à une composante morpholexicale la génération des représentations phonologiques des formes fléchies. Une méthode simple pour parvenir à ce résultat est de créer un lexique de formes fléchies phonologiques.

Nous proposons un modèle de composante phonologique [Pérennou,1992b] qui permet de transcrire la représentation phonologique issue de la composante morpholexicale. Cette composante phonologique est caractérisée par :

- l'utilisation des groupes phonologiques contextuels dans la représentation phonologique et la forme des règles phonologiques qui leur sont associées,
- la représentation de sortie du traitement, appelée transcription phonotypique, qui vise à être un modèle probabiliste des prononciations possibles dans le dialecte choisi,
- l'utilisation des groupes à prononciations multiples comme cellules de base de la transcription phonotypique,
- la stratégie de transcription qui procède, pour chaque point de l'entrée, en une seule étape, de sorte qu'il n'y a aucune représentation intermédiaire entre l'entrée phonologique et la représentation phonotypique.

2.1. La représentation phonologique

La représentation phonologique de chaque mot fléchi est produite de manière à rendre compte des variantes possibles de prononciation dans la population concernée par l'application. Pour cela, nous allons introduire les groupes à prononciations multiples (ou gpm) et les groupes phonologiques contextuels (ou gpc).

Les gpm rendent compte de toutes les variantes possibles de prononciation d'un ou de plusieurs phonèmes contigus pour un groupe de locuteurs donné. Ils permettent de traduire des phénomènes phonologiques comme la chute du schwa ou les assimilations...

Les gpc rendent compte des variantes de prononciations dues au contexte du mot dans la phrase. Ainsi les variations de prononciation contextuelles sont prises en compte dès les représentations lexicales.

Pour représenter les gpc et les gpm, nous adopterons les conventions suivantes :

- si la sous-chaîne phonétique x a une, et une seule, prononciation, ses représentations phonologique et phonétique seront x (i.e. x est un gpc trivial et un gpm trivial).
- si x est constitué d'un ou plusieurs phonèmes interdépendants dont les réalisations sont variables selon le contexte alors <x> est un gpc non trivial.
- si ses prononciations multiples ne dépendent pas du contexte alors (x) est un gpm non trivial.

On utilisera dans la suite les symboles API sauf si des codes spéciaux sont nécessaires pour les gpc et gpm. Par exemple, le mot *membre* est représenté par "mɑ̃<~brə>" au niveau phonologique. Le gpc <~brə> engendre le gpm (~brə>) devant une pause ou le gpm (~brə) devant une consonne ou /br/ devant une voyelle ou une semi-voyelle. Les gpm apparaissent au niveau phonotypique. Pour des exemples plus détaillés, on se reportera à [de Ginestel-Mailland,1993a].

Les gpm peuvent être probabilisés. Ainsi, la règle phonétique pour le gpm (~brə) qui associe les réalisations brŒ, b et m est représentée par :
(~brə) : brŒ 0.6 , b 0.3 , m 0.1 (1)
où Œ est un gpm trivial ou non selon la finesse du modèle phonologique. En effet, pour un modèle très détaillé le gpm Œ pourra avoir les réalisations ø, ø̞ ou œ (où ø̞ est un son intermédiaire entre ø et œ) alors que dans la version simplifiée du modèle nous ne distinguons pas ce degré de finesse.

Ainsi définis les gpm peuvent être associés de diverses manières à des réseaux probabilistes, en particulier à des HMM.

Le modèle utilisé ici est simplifié. Nous ne nous intéressons qu'aux gpm finaux qui décrivent les phénomènes phonologiques de liaison. Quelques gpm non triviaux internes sont conservés pour les mots d'origine étrangère.

2.2. La composante phonologique

Le but de notre composante phonologique est de transformer une forme orthographique en une forme phonotypique. Cette unique transcription rend compte de toutes les variantes de prononciation.

Le mécanisme de génération [de Ginestel-Mailland,1993b] est composé de trois modules :
• Le module d'accès lexical qui génére la forme phonologique avec catégories syntaxiques à partir de la forme orthographique. La forme phonologique est composée de gpc.
• Le module de placement de frontières qui enrichit la forme phonologique de frontières phonologiques.
• Le module de génération phonologique qui produit la transcription phonotypique à partir de règles phonologiques. Une forme phonotypique est constituée de gpm.

Pour chaque mot d'un énoncé, un lexique dérivé de BDLEX fournit la représentation phonologique et la classe syntaxique [Pérennou,1992a].

D'après [Delattre,1966], les frontières entre les mots sont placées en fonction des catégories syntaxiques du mot précédent, du mot courant et du mot suivant dans un modèle tri-classe. Dans notre modèle simplifié, les frontières sont placées uniquement selon la catégorie syntaxique du mot courant.

Pour générer la forme phonotypique de l'énoncé, nous utilisons un ensemble de règles phonologiques définies dans [Pérennou,1992b] ainsi que les symboles de frontières décrivant les phénomènes suivants :

: liaison obligatoire
| : pause
? : liaison variable.

Un exemple de génération de forme phonotypique est donné figure 1. Les catégories syntaxiques sont celles de BDLEX (J : adjectif, d : déterminant, N : nom, V : verbe,...).

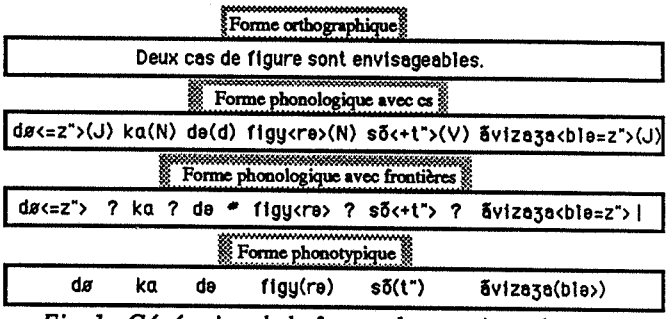


Fig. 1 : Génération de la forme phonotypique de l'énoncé "Deux cas de figure sont envisageables."

3. L'APPRENTISSAGE

Le modèle d'apprentissage de règles phonologiques et phonétiques que nous proposons fait appel à une variante du système d'alignement VERIPHONE décrit dans [Pérennou,1989]. Il prend en entrée une chaîne de gpm qui est la transcription phonotypique issue de la composante phonologique décrite précédemment et une chaîne qui est la transcription phonétique provenant par exemple d'un corpus de parole. Le système d'alignement assigne à chaque gpm une suite d'unités phonétiques par application de règles phonétiques. Le système produit en sortie l'énoncé aligné sous forme d'une liste de couples gpm/phonèmes.

Les règles phonétiques et les probabilités associées aux réalisations phonétiques sont données à l'initialisation du système par un expert. Celles-ci sont ajustées et adaptées par le module d'apprentissage. Les fréquences associées à chaque réalisation de gpm sont incrémentées à chacune de leurs apparitions. On peut

alors procéder par estimation comme dans le cas des HMM.

L'apprentissage des règles phonétiques exige des corpus transcrits phonétiquement. La taille du corpus doit être suffisante pour que tous les gpm y soient attestés plusieurs fois de manière à permettre des estimations statistiques correctes. Ces statistiques peuvent être combinées avec des estimations a priori fournies par des experts phonéticiens.

Dans le cadre du Pôle Parole du GDR-PRC CHM, nous avons entrepris de mettre en œuvre cette méthodologie sur le corpus BREF80. L'étude de la répartition statistique des triphones dans ce corpus, effectuée par le laboratoire LIMSI [Lamel, 1991] montre qu'il offre une base d'investigation intéressante.

La figure 2 montre le système que nous mettons en œuvre pour l'apprentissage des règles phonétiques.

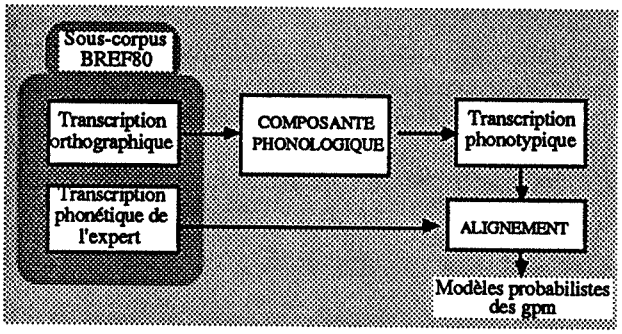


Fig. 2 : Apprentissage des règles phonétiques.

4. RESULTATS

Nous allons maintenant donner un ensemble de résultats obtenus sur une partie du corpus BREF80 composée de 1451 phrases distinctes et 80 locuteurs. Chaque locuteur a prononcé en moyenne 23 phrases. Celles-ci sont de deux types :

a) phrase contenant presque tous les phonèmes

b) phrase prononcée en contexte (15 mots par phrase).

Ce corpus a été transcrit par le LIMSI en 2 étapes :

a) le système de phonétisation graphème-son GRAPHON développé au LIMSI a fourni une transcription phonétique à partir de la transcription orthographique des textes

b) la transcription phonétique a ensuite été corrigée par des experts après écoute du signal.

Cette transcription n'est pas exempte de reproches. Les assimilations de voisement et de nasalité en particulier ont été la plupart du temps ignorées. L'apprentissage que nous avons donc fait sur ce corpus concerne donc la prononciation effective filtrée par GRAPHON et les transcripteurs.

Dans une étape future, il faudra remédier à ces défauts. Cependant, il nous a paru intéressant de montrer déjà des résultats sur la liaison, le schwa et les liquides dans les groupes finaux

4.1. Fréquence des GPM dans le corpus

Dans le sous-corpus de BREF80 que nous avons utilisé, 291 gpm distincts ont été recensés. Le nombre d'occurences de ces gpm est de plus de 120.000. La figure 3 donne leur répartition.

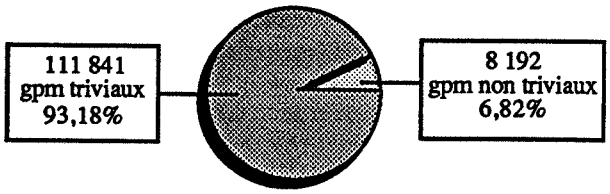


Fig. 3 : Répartition des gpm

Les gpm non triviaux étudiés ici sont en finale de mot ou internes pour des mots étrangers. Si le nombre de gpm non triviaux est peu élevé, nous pouvons penser que l'étude des gpm internes augmentera ce nombre significativement. En effet, l'ajout des gpm internes sur les séquences Ce ou CLe donnera un espace de recherche plus étendu.

Les gpm triviaux sont en grande partie représentés par les phonèmes courants. Leur répartition en classe est donnée dans la figure 4.

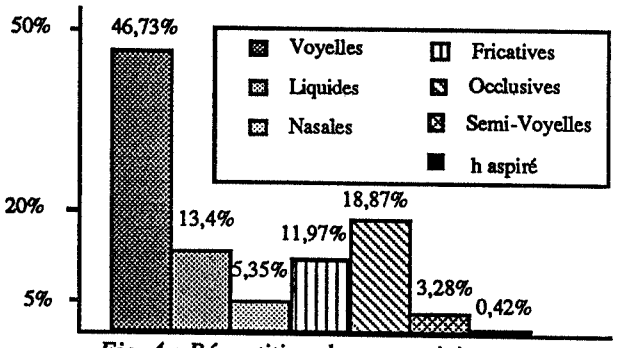


Fig. 4 : Répartition des gpm triviaux

La distribution des gpm non triviaux est donnée dans la figure 5. Seuls les gpm suffisamment représentés sont indiqués. On note les ensembles :

C	consonnes
O	obstruantes
P	plosives
F	fricatives
M	sonantes
L	liquides
N	nasales

et les gpm :

(...>) : gpm avant une pause

(C") : consonne latente

(...z") : gpm de mots au pluriel

(...t") : gpm de verbes à la troisième personne du pluriel

(^...) : assimilation de voisement
(~...) : assimilation de nasalité.
Lorsque un 'e' ne se trouve pas dans la graphie du mot mais peut être éventuellement inséré, on le note @.

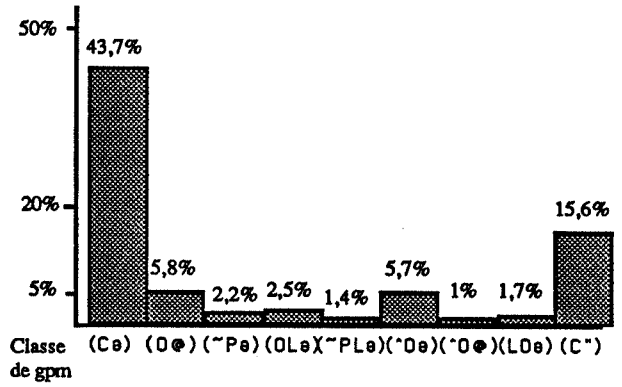


Fig. 5: Répartition des gpm non triviaux

Les deux classes les plus représentées correspondent à presque 60% des occurrences. Celles-ci caractérisent des gpm simples, alors que les gpm plus complexes (composés de 2 consonnes) : (LOə), (~PLə) et (OLə) ne dépassent pas 6% des cas.

4.2. Les prononciations des principaux gpm

Un exemple de représentation des règles phonétiques est donné en (1). Les réalisations possibles d'un gpm décrivent l'introduction ou non d'un schwa dans une liaison, la troncation d'un /ə/ et d'une (ou de plusieurs) consonne(s) avant une pause, la nasalisation après une voyelle nasale, etc.

Toutes les constatations que nous allons faire ici ne sont pas exhaustives. L'étude de tout le corpus de BREF80 nous donnera des résultats plus complets et certains gpm qui ont très peu d'occurences sur ce sous-corpus pourront l'objet de discussions détaillées.

4.2.1 Réalisations des gpm (Cə) et (C@)

La figure 6 montre par exemple la répartition des réalisations pour les gpm de type (Cə). En moyenne, dans 88,9% des cas le /ə/ est tronqué.

La figure 7 nous donne la répartition des réalisations du gpm (P@). Ce gpm représente une plosive en fin de mot après laquelle un schwa a pu être inséré par exemple dans la séquence "bec camus" (-> bɛ(k@) kamy).

Gpm	Réalisations			
(Lə)	LŒ	7%	L	93%
(Nə)	NŒ	15,6%	N	84,4%
(Fə)	FŒ	9,4%	F	90,6%
(Pə)	PŒ	12,3%	P	87,7%

Fig. 6 : Réalisations des gpm (Cə)

Gpm	Réalisations			
(P@)	PŒ	7,4%	P	92,6%

Fig 7 : Réalisations du gpm (P@)

La comparaison des réalisations des gpm (Pə) et (P@) et leur répartition nous permettent de penser qu'il n'est peut-être pas nécessaire de différencier ces deux gpm ; autrement dit, qu'une plosive soit suivie ou non d'un 'e' dans la graphie, une fréquence analogue des réalisations PŒ est observée.

4.2.2 Réalisations des gpm (LOə) et (LO@)

La figure 8 représente la répartition des réalisations de gpm plus complexes (LOə) et (LP@). Le phénomène observé plus haut pour (Pə) et (P@) ne se renouvelle pas ici car les fréquences d'apparition des gpm (LPə) et (LP@) sont très différentes. Curieusement, ce sont les fréquences de (LFə) et (LP@) qui paraissent être plus ressemblantes.

Gpm	Réalisations			
(LPə)	LPŒ	60,7%	LP	39,3%
(LFə)	LFŒ	37%	LF	63%
(LP@)	LPŒ	31,6%	LP	68,4%

Fig.8 : Réalisations des gpm (LOə) et (LP@)

4.2.3 Réalisations des gpm (OLə)

La figure 9 représente les résultats obtenus pour les gpm (OLə).

Gpm	Réalisations			
(Olə)	OLŒ	97%	O	3%

Fig.9 : Réalisations des gpm (OLə)

Ces fréquences nous montrent que dans un corpus lu, les gpm (OLə) sont prononcés en extension. Ce phénomène est beaucoup plus marqué que dans les autres cas.

4.2.4 Réalisations des gpm en C"

La fréquence de réalisation des liaisons obligatoires dépend de nombreux facteurs : stylistique, syntaxique, prosodique, historique, ...[Delattre, 1966], [Encrevé, 1988]. Nous allons maintenant isoler les cas de liaison pour un nom au pluriel et un verbe à la troisième personne du pluriel se terminant par "ent". D'après P. Delattre, «la liaison se fait davantage au pluriel (...) pour distinguer le pluriel du singulier». La consonne de liaison facultative notée (C") peut quelquefois être associée aux consonnes et au schwa la précédant pour former un gpm plus complexe : (CəC") ou (CCəC").

La figure 10 montre les fréquences de liaison pour certaines classes de gpm finaux de noms au pluriel. Nous pouvons citer les exemples suivants issus du corpus pour illustrer ces classes :

"...dérapages individuels, ..."
-> derapa(ʒəz") ẽdividuel
"...parlementaires américains, ..."
-> parlemãtɛ(rəz") amerikẽ
"...provinces iraniennes, ..."
-> prɔvẽ(^səz") iranjɛnə
"...journalistes occidentaux, ..."
-> ʒurnali(stəz") ɔksidãto

Classe de gpm	Liaison	Non-liaison
(Oəz")	20%	80%
(Məz")	31,4%	68,6%
(^Oəz")	14,3%	85,7%
(CCəz")	7,4%	92,6%

Fig. 10 : Fréquences des liaison après un mot au pluriel

Les liaisons les plus fréquentes se font pour la classe (Məz") sans pourtant atteindre un taux supérieur aux non-liaisons. En moyenne, la fréquence de liaison pour ces classes est de seulement 18,27%. La liaison paraît se faire d'autant plus facilement que le groupe final (ou ici la syllabe finale) est simple (ie qu'il est composé de peu de consonnes). En effet, le minimum d'enchaînements se situe dans la classe complexe (CCəz").

La figure 11 donne les fréquences de liaison après un verbe à la troisième personne du pluriel. Les séquences suivantes sont des exemples de ces gpm :
"...participent au ..."
-> partisi(pət") o
"...prennent un ..."
-> prɛ(nət") ẽ
"...regardent aux ..."
-> rəgə(rdət") o

Classe de gpm	Liaison	Non-liaison
(Oət")	20%	80%
(Mət")	27,3%	72,7%
(CCət")	20%	80%

Fig. 11 : Fréquences des liaisons après un verbe à la troisième personne du pluriel

Nous retrouvons dans ce cas le maximum de liaison pour un groupe final en sonante. La moyenne des fréquences de liaison est ici plus élevée : 22,43%.

La liaison paraît donc se faire plus facilement après un verbe qu'après un nom et plus fréquemment pour un pluriel en sonante.

5. CONCLUSION

Nous avons montré qu'en adoptant le modèle phonologique basé sur les groupes à prononciations multiples (gpm) et les groupes phonologiques contextuels (gpc), il était possible d'apprendre automatiquement les probabilités de prononciation associées aux gpm. Ceci a été rendu possible grâce à un lexique phonologique où les mots sont représentés en gpc et à deux bases de règles. La première définit une fois pour toutes les gpc. La seconde décrit les gpm et est adaptable par apprentissage à partir d'un corpus transcrit.

Les résultats que nous avons obtenus actuellement sur une partie du corpus BREF sont encore partiels. Ils concernent un groupe de locuteurs dans une situation de communication donnée — à savoir lecture de documents. De tels résultats montrent simplement une possibilité d'apprentissage pour une tâche déterminée de communication orale homme-machine.

Ces résultats permettent déjà de se faire une bonne idée de l'impact des variations phonologiques dans les productions orales et par là même de mieux situer le rôle de la composante phonologique dans les systèmes de reconnaissance de la parole.

6. REFERENCES

[Delattre,1966] P. Delattre, "La liaison en français - Tendances et classification", "La fréquence des liaisons facultatives en français", "Les facteurs de la liaison facultative en français", Studies in French and Comparative Phonetics, Mouton, The Hague,1966.

[Encrevé,1988] P. Encrevé, "La liaison avec et sans enchaînement- Phonologie tridimensionnelle du français", Le Seuil.

[de Ginestel-Mailland,1993a] A. de Ginestel-Mailland, G. Pérennou, M. de Calmès, "Une approche de la phonologie en reconnaissance de la parole", 2èmes journées internationales "L'interface des mondes réels et virtuels", Montpellier, 22-26 mars1993, pp. 345-354.

[de Ginestel-Mailland,1993b] A. de Ginestel-Mailland, M. de Calmès, G. Pérennou, "Multi-Level Transcription of Speech Corpora from Ortographic Form", Proceeding of EUROSPEECH93, Berlin, september 1993.

[Lamel,1991] L. F. Lamel, J.-L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French", Eurospeech 91, Genova, 24-26 Septembre 1991.

[Pérennou,1989] G. Pérennou, H. Kabré, M. de Calmès, J.-M. Pécatte, N. Vigouroux, "Une approche de l'étiquetage automatique indépendant du locuteur", Séminaire variabilité du locuteur, Avignon, 20-21 Juin 1989.

[Pérennou,1992a] G. Pérennou, D. Cotto, M. de Calmès, I. Ferrané, J.-M. Pécatte, "Le projet BDLEX de base de données lexicales du français ecrit et parlé", Séminaire lexique, IRIT-UPS, Toulouse, 21-22 Janvier 1992.

[Pérennou,1992b] G. Pérennou, "Introduction aux groupes à prononciations multiples", Rapport Interne IRIT.

CATÉGORISATION PHONÉTIQUE ET SUBPHONÉTIQUE DANS UN LEXIQUE DU FRANÇAIS
ÉTUDE PRÉLIMINAIRE

Hocine AIT-MESSAOUD, Véronique AUBERGÉ, Louis-Jean BOË

INSTITUT DE LA COMMUNICATION PARLÉE
CNRS URA 368, INPG/Univ. Stendhal, BP 25 38040 Grenoble cedex 9
(hocine, auberge, boe)@icp.grenet.fr

Résumé

Ce travail fait partie d'un vaste programme dont le but est de rassembler des matériaux lexicaux et de les intégrer dans un large lexique avec plusieurs niveaux de représentation (orthographiques, phonétiques et subphonétiques). Ce projet est mené en collaboration avec différents laboratoires français de traitement de la langue naturelle et de la parole (Groupe de Travail *Lexique* du PRC Informatique). Nous présentons les premiers résultats d'une étude sur l'organisation des cohortes (CV) et la structuration subphonétique d'un sous-lexique. Première application visée : le filtrage lexical en reconnaissance de la parole.

Mot-clés: *lexiques, cohortes, catégorisation subphonétique, macro-traits.*

1. INTRODUCTION

Ce projet s'inscrit dans un axe de recherche général qui a pour but de tester, avec des matériaux orthographiques et phonétiques[1] suffisamment représentatifs, des hypothèses sur l'émergence du code (segments, syllabes...), en particulier la structuration de séquences des sons dans les lexiques (cf. par exemple [Maddieson, 1993]). Il se donne pour moyen de construire, structurer, gérer et enrichir un lexique multi-niveaux du français.

Cette étude préliminaire est centrée sur les cohortes CV, le « squelette syllabique » du mot et des macro-traits issus de l'expertise et utilisés dans un module de décodage acoustico-phonétique d'un système de reconnaissance de la parole [Méloni, 1982].

Cette recherche est soutenue par des actions nationales :

- Les projets « Filtrage lexical » et ORALL[1] (Organisation Représentation et Accès à de Larges Lexiques) [Aubergé et al., 1993] du GDR-PRC Communication Homme-Machine.
- Le Groupe de Travail (GT5) « Lexique » du PRC Informatique[2].

La base de données ORALL se présente comme un dépôt de connaissances issues des différents niveaux de représentations de la parole. C'est l'intégration commune de ces connaissances qui va permettre, en retour, de gérer de nouvelles informations.

Plusieurs matériaux sont donc à l'origine de ORALL. L'un d'entre eux est un lexique du français qui réunit environ 60 000 formes orthographiques canoniques (« Le 60 000 » de l'ICP) et leurs catégories lexicales. Chaque forme orthographique est associée à une forme phonétique de référence choisie dans le Petit Robert 1. Une autre source est le corpus BDPHO de vingt heures de parole spontanée transcrite phonétiquement [Boë et Tubach, 1992a]. Le lexique « De A à Zut » [Boë et Tubach, 1992b] extrait de BDPHO et lemmatisé [Dujardin et al., 1994] contient les fréquences et les formes des variantes observées. Une autre référence est la base de données BRULEX développée au Laboratoire de Psychologie Expérimentale de Bruxelles [Content, 1990].

Nous présentons d'abord les différentes fonctionnalités de ORALL mais nous nous intéressons plus particulièrement aux fonctions de catégorisation phonétique (cohortes CV) et subphonétique (macro-traits). À chaque élément de la structure cohorte est attribuée la liste des macro-traits qui le définit. Des résultats statistiques préliminaires sur la typologie des structures cohortes/macro-traits, sur un sous-lexique d'environ 50 000 formes canoniques, sont finalement proposés.

2. FONCTIONNALITÉS DU LEXIQUE

Le lexique ORALL se présente comme un ensemble de champs : formes orthographiques, formes

[1] Ce projet, coordonné par l'ICP, a rassemblé plusieurs laboratoires français : le laboratoire Parole et Langage (Aix-en-Provence), le Laboratoire d'Informatique d'Avignon, l'équipe LGI-TRILAN (Grenoble), l'URA 1033 (Paris X) et la société SITE (Paris). Il se différencie du projet BDLEX orienté phonologie [Pérennou, 1988].

[2] Il est constitué des membres d'ORALL et de laboratoires de Traitement de Langue Naturelle 2LC (Nice), CRIN-LORIA (Nancy), CRISSTAL (Grenoble), GETA (Grenoble), CNRS-HESO (Paris), IPA (Aix-en-Provence), IRIN (Nantes), LGI-TRILAN (Grenoble), LITP (ENS Fontenay St Cloud), ICP (Grenoble) et LIA (Avignon) auxquels se joignent Dpt de Linguistique (Toulouse Le Mirail), LPG (Genève), ERSS et IRIT (Toulouse).

phonétiques de référence, variantes observées dans BDPHO, variantes attestées, catégories lexicales, structures cohortes, macro-traits, points d'unicité, voisinages phonétiques et orthographiques. Les champs cohortes et macro-traits sont définis par une fonction de génération. Le choix de représentation de ces champs, soit par extension (liste des items), soit par intention (fonction associée), dépend de l'application visée. Dans le cas du filtrage lexical, le but est de générer un sous-lexique qui contient explicitement, et en accès direct, les champs orthographique, phonétique, cohortes et macro-traits.

Un environnement de travail standardisé et convivial (ORALL s'adresse aussi à des lexicologues, phonéticiens, psycholinguistes...) a été développé autour du lexique. Le choix du Macintosh, du langage de programmation HyperTalk sous HyperCard pour l'interface, associé au langage *Think C* pour le gestionnaire, favorise l'ouverture des traitements à des utilisateurs non informaticiens. Les formes phonétiques sont représentées par les caractères de l'Alphabet Phonétique International (API).

2.1. Cohortes et macro-traits

Nous avons privilégié une représentation en classes *Consonne Voyelle* qui renvoie à une modulation première des gestes de production (de la fermeture à l'ouverture du conduit vocal) [Brodda, 1979 ; Lindblom, 1983]. Ces cohortes font appel à des catégorisations linguistiques [Halle et Vergnaud, 1980]. Elles ont déjà figurées dans BDLEX dès 1986 [Pérennou et Calmès, 1986]. Plus généralement, la notion de cohorte a été utilisée dans des modèles d'accès lexical [Marslen-Wilson, 1984, 87]. Les cohortes constituent un matériau de base relativement peu exploité en synthèse et en reconnaissance de la parole [Carlson et al., 1985].

Les cohortes ainsi dérivées seront structurées dans un deuxième temps en unité rythmique minimale (syllabe). On adoptera la décomposition classique de la syllabe : *attaque, noyau vocalique et coda,* dans laquelle l'attaque et la coda peuvent être réduites à la chaîne vide [Aubergé et al., 1988].

Au niveau segmental l'analyse permet d'extraire, de la redondance, les traits distinctifs phonologiques. Depuis Jakobson, Fant et Halle [Jakobson et al., 1952], les unités distinctives ont fait l'objet de plusieurs catégorisations phonologiques. Mais celles-ci se prêtent mal, à cause des phénomènes de coarticulation, à une implantation immédiate, que ce soit en synthèse ou reconnaissance de la parole. Depuis les travaux de Wiren et Stubbs [Wiren et Stubbs, 1956], ont été affinés des « macro-traits » acoustico-phonétiques, plus robustes, mais aussi plus génériques [Caelen, 1979 ; Méloni, 1982 ; De Mori, 1983 ; Mercier, 1983 ; Rossi et al., 1983 ; Bulot, 1987]. Nous avons retenu les traits utilisés par le Laboratoire d'Informatique d'Avignon, dans leur dernière version mise au point par [Gilles, 1993] (Liste 1).

fricatif	: 1	grave	: 11
non-fricatif	: 2	aigu	: 12
occlusif	: 3	compact	: 13
non-occlusif	: 4	diésé	: 14
voisé	: 5	fermé	: 15
non-voisé	: 6	ouvert	: 16

Liste 1 : Codage des macro-traits

Les macro-traits retenus pour les voyelles (Liste 2) et les consonnes (Liste 3) sont représentés ci-dessous.

Voyelles	grave	aigu	compact	diésé	fermé	ouvert
i	–		–		+	
e, y	–		–	–	+	
ɛ	–			–		
o		–		–	+	
ɔ		–		–		+
u		–	–	–	+	
ø				–	+	
ɑ̃				–		
ɔ̃		–		–		
œ, œ̃, ɛ̃, a				–		+

+ : obligation d'avoir le trait correspondant.
– : interdiction d'avoir le trait correspondant.

Liste 2 : Les macro-traits pour les voyelles.

Consonnes	occlusif	fricatif	voisé
p, t, k	+	–	–
b, d, g		–	+
f		+	–
s, ʃ	–	+	–
v, ʀ	–		
z, ʒ	–	+	
m, n, ŋ, ɲ, l, w, j, ɥ	–	–	+

+ : obligation d'avoir le trait correspondant.
– : obligation d'avoir le trait négatif (c'est-à-dire Non-Fricatif, Non-Occlusif ou Non-Voisé).

Liste 3 : Les macro-traits pour les consonns.

Afin d'optimiser le traitement et l'espace mémoire, nous avons codé numériquement chaque macro-trait. Les trente sept sons adoptés pour les formes phonotypiques du lexique – vingt deux consonnes et semi-consonnes, onze voyelles orales et quatre voyelles nasales – sont donc décrits par ces macro-traits, à l'exception du *schwa* [ə] et du *h* aspiré ['], qui ont fait l'objet d'un codage spécial.

Des traitements ont été développés en vue de réaliser différentes tâches nécessaires pour coder et procéder à des analyses statistiques sur les cohortes. Cet ensemble de fonctionnalités s'articule autour d'un gestionnaire superviseur [Ait-Messaoud, 1993]. Les résultats intermédiaires de toutes les tâches réalisées par ce gestionnaire sont rangés dans des fichiers ASCII. L'une des fonctions permet de faire la conversion de l'entrée phonétique en cohortes et en macro-traits.

La complexité de cette fonction est linéaire puisque les entrées sont converties séquentiellement (Liste 4). L'utilisation des codes permet de limiter la taille du fichier résultat, sachant que ce dernier est presque dix fois plus grand que le fichier des entrées (le champ phonétique occupe une taille d'environ 500 Ko).

```
bajəmɑ̃ CVCVCV 2 5, 14 16, 2 4 5, 11 11 11, 2 4 5, 14
baje CVCV 2 5, 14 16, 2 4 5, 11 13 14 15
bajœʀ CVCVC 2 5, 14 16, 2 4 5, 14 16, 4
baji CVCV 2 5, 14 16, 2 4 5, 11 13 15
bajɔ̃ CVCV 2 5, 14 16, 2 4 5, 12 14
baiʀam CVVCVC 2 5, 14 16, 11 13 15, 4, 14 16, 2 4 5
bɛse CVCV 2 5, 11 14, 1 4 6, 11 13 14 15
bɛsje CVCCV 2 5, 11 14, 1 4 6, 2 4 5, 11 13 14 15
baʒu CVCV 2 5, 14 16, 1 4, 12 13 14 15
balata CVCVCV 2 5, 14 16, 2 4 5, 14 16, 2 3 6, 14 16
balɛs CVCVC 2 5, 14 16, 2 4 5, 11 14, 1 4 6
balizœʀ CVCVCVC 2 5, 14 16, 2 4 5, 11 13 15, 1 4, 14 16, 4
balast CVCVCC 2 5, 14 16, 2 4 5, 14 16, 1 4 6, 2 3 6
bale CVCV 2 5, 14 16, 2 4 5, 11 13 14 15
balɔ̃ CVCV 2 5, 14 16, 2 4 5, 12 14
```

Liste 4 : Un extrait de l'entrée phonétique avec les cohortes et macro-traits associés

Un tri sur les cohortes complète les fonctionnalités du gestionnaire. Son principe est simple : on recherche à chaque passage le minimum et on le range dans un nouveau fichier. Ce minimum est déterminé par l'ordre alphabétique et par la longueur de la cohorte, ce qui permet de regrouper les cohortes de même longueur et de les ordonner. Le traitement est de complexité $O(n^2)$, si n est la taille du fichier contenant les cohortes.

Un autre traitement permet d'extraire n'importe quelle classe de cohortes contenue dans un lexique phonétique converti en cohorte et en macro-traits [Ait-Messaoud, 1993]. Le fichier résultat (Liste 5) contient, pour chaque classe de cohortes, les occurrences (en première colonne) de toutes les classes de macro-traits :

```
2       CVCV 2 5, 14 16, 2 4 5, 11 13 14 15
1       CVCV 2 5, 14 16, 2 4 5, 11 13 15
2       CVCV 2 5, 14 16, 2 4 5, 12 14
1       CVCV 2 5, 11 14, 1 4 6, 11 13 14 15
1       CVCV 2 5, 14 16, 1 4, 12 13 14 15
```

Liste 5 : Un extrait de cohortes et leurs occurrences

Les différents traitements du gestionnaire, présentés ci-dessus, privilégient plutôt l'optimisation de l'espace mémoire par rapport au temps de calcul. L'ensemble des traitements a été développé compte tenu du choix informatique (micro-calculateur).

2.2. Voisinage et points d'unicité

Le point d'unicité d'un mot peut être défini, par rapport à un ensemble donné de mots (le lexique), comme la sous-chaîne plus longue d'un caractère au plus long préfixe commun à l'ensemble des mots. Deux mots sont donc voisins lorsqu'ils ne diffèrent que d'un caractère, tous les autres étant identiques et dans le même ordre. Cette définition peut être généralisée aux chaînes cohortes : dans ce cas, le point d'unicité correspond à un nœud d'un arbre si on considère que le lexique est organisé sous forme d'arborescences. La racine correspond à la classe d'éléments communs au sous-lexique et le parcours de l'arbre permet d'accéder à la cohorte ou au groupe de cohortes ayant des macro-traits identiques. La connaissance du point d'unicité d'un lexique ou sous-lexique donné peut intervenir dans une euristique du filtrage lexical. L'ensemble des voisins d'un mot est défini globalement comme les mots du lexique obtenus par insertion, effacement ou substitution d'un caractère phonétique ou orthographique. C'est donc l'opération de base d'un accès tolérant au lexique. Ces deux fonctions, voisinage et point d'unicité, sont des notions inhérentes à la plupart des modèles psycholinguistiques de reconnaissance (identification et accès) de mots [Kolinsky et al., 1991].

3. PREMIERS RÉSULTATS

Dans ce qui suit, nous présentons les résultats obtenus après traitements sur les structures cohortes. Une sélection du lexique constituée de 48 172 entrées est utilisée (sous-lexique du « Le 60 000 »). Parmi les cohortes extraites figurent 1 094 cohortes distinctes, et 15 d'entre elles permettent de couvrir 50% de l'ensemble des occurrences. Le début de la liste des cohortes par occurrences est présenté (Liste 6) avec sa loi de distribution (Figure 1) : on peut noter la régularité de Zipf [1935].

occurrences	**cohortes**	**%**
4493	• CVCVCV	9,33
3115	• CVCV	6,47
2689	CVCVCVCV	5,58
1898	CVCCVCV	3,94
1711	CVCCV	3,55
1426	CVCVCCV	2,96
1276	• CCVCV	2,65
1185	CCVCVCV	2,46
1113	• VCVCV	2,31
1103	CVCVC	2,29
1094	VCVCVCV	2,27
1025	CVCCVCVCV	2,13
801	CVCVCCVCV	1,66
795	CVCVCVC	1,65
789	CVCVCVCCV	1,64
24513	**sous total**	**> 50%**
747	CVCVCVCVCV	1,55
730	VCCVCV	1,52
590	CVCCVC	1,22
582	CCVCVCVCV	1,21
492	VCCVCVCV	1,02
481	VCVCCV	1,00
477	CVCCVCCV	0,99
400	VCVCCVCV	0,83
29407	**total**	**> 60%**

Liste 6 : Les 25 premières cohortes représentant 60% du total

Remarquons l'écrasante majorité des syllabes CV qui, dans le lexique du français comme dans toutes les langues, représente une tendance universelle [Maddieson et Precoda, 1992].

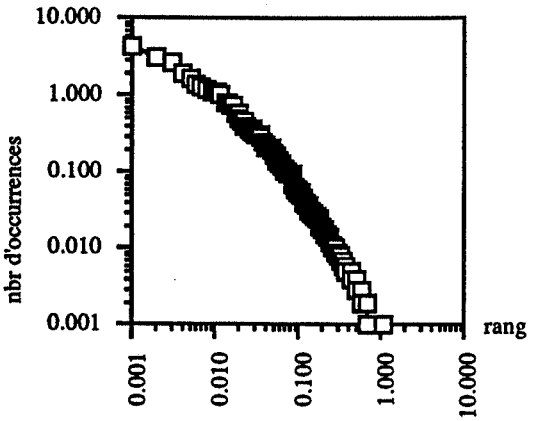


Figure 1 : Distribution rang-occurrences des cohortes

Nous détaillons plus loin les résultats des cohortes les plus occurrentes : CVCVCV, CVCV, CCVCV dont l'attaque est un groupe consonantique et la cohorte VCVCV, la plus occurrente des cohortes à attaque vocalique.

3.1. Cohortes CVCVCV

Leur fréquence représente un peu plus de 9% (4 493 occurrences) du total des cohortes. Si on distingue les représentants de la cohorte CVCVCV selon les classes définies par les macro-traits, on obtient 3 568 classes distinctes, soit 79% des cohortes. La distribution rang-occurrences (Figure 2) est quasi-linéaire sur une échelle Log (Log).

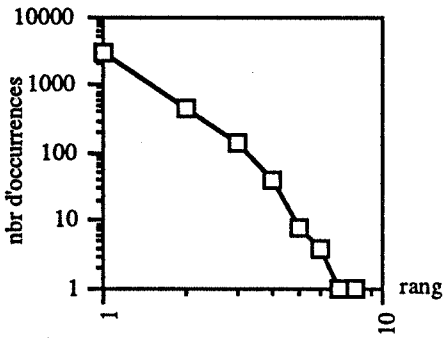


Figure 2 : Distribution rang-occurrences des cohortes CVCVCV

Parmi les cohortes les plus fréquentes (Liste 7) on remarque que les sons les plus ambigus présentent les macro-traits suivants :

- occlusif, non-fricatif, non-voisé,
- non-fricatif, non-occlusif, voisé,
- non-occlusif,
- fricatif, non-occlusif, non-voisé,
- fricatif, non-occlusif,

et à un degré moindre : occlusifs voisés,
et pour les noyaux vocaliques :

- grave, compact, diésé, fermé,
- diésé,
- aigu, compact, diésé, fermé,
- grave, diésé,
- diésé, ouvert,
- le schwa en position finale,

et à un degré moindre pour le noyaux vocaliques :

- aigu, diésé, fermé,
- aigu, diésé, ouvert,
- grave, compact, fermé,

Inversement, pour les attaques syllabiques, n'ont pas un degré d'ambiguïté élevé les sons ayant les macro-traits suivants :

- fricatif, non-voisé,

et pour les noyaux vocaliques :

- aigu, diésé.
- diésé, fermé.

8 [p t k][a ɛ œ̈ œ][j ɥ w l m n ŋ ɲ][a ɛ œ̈ œ][z ʒ][ə]
7 [j w l m n][a ɛ œ̈ œ][p t k][a ɛ œ̈ œ][z ʒ][ə]
6 [p t k][a ɛ œ̈ œ][j ɥ w l m n ɲ ŋ][ɛ][p t k][ə]
6 [b d g][e y][b d g][u][p t k][e y]
6 [j w l m n ɲ][i][j ɥ w l m n ŋ ɲ][a ɛ œ̈ œ][z ʒ][ə]
6 [j w l m n ɲ][a ɛ œ̈ œ][p t k][i][j ɥ w l m n ŋ ɲ][ye]
5 [ʀ v][e y][p t k][e y][j ɥ w l m n ŋ ɲ][e y]
5 [j w l m n ɲ][e y][p t k][i][j ɥ w l m n ŋ ɲ][e y]
5 [b d g][e y][p t k][a ɛ œ̈ œ][j ɥ w l m n ŋ ɲ][e y]
5 [p t k][ɔ][j ɥ w l m n ŋ ɲ][ɛ][ʀ v][ə]
5 [p t k][ɑ̃][p t k][a ɛ œ̈ œ][j ɥ w l m n ŋ ɲ][ə]
5 [b d g][e y][p t k][ɔ][j ɥ w l m n ŋ ɲ][e y]
5 [p t k][a ɛ œ̈ œ][j ɥ w l m n ŋ ɲ][a ɛ œ̈ œ][s ʃ][ə]
5 [p t k][a ɛ œ̈ œ][j ɥ w l m n ŋ ɲ][i][s ʃ][ə]

Liste 7 : Les occurrences les plus fréquentes des cohortes CVCVCVde macro-traits identiques et les sons possibles

Les plus fortes ambiguïtés concernent [p t k], [b d g] et la macro classe des semi-voyelles, liquide et nasales [j w l m n ɲ], ([ɥ, ŋ] n'apparaissant pas en position initiale). Remarquons au passage que le triplet d'occlusives sourdes est présent dans 92% des langues [Bernathova et Zackova, 1992]. Pour les voyelles, il faudra séparer [e y] et les voyelles ouvertes [a ɛ œ̈ œ].

La typologie par cohortes met bien en évidence la structure phonologique de la langue [Togeby, 1965], comme cela a été vérifié en particulier pour les groupes consonantiques [Aubergé, 1991]. Si le filtrage permet de segmenter le lexique, il regroupe aussi des catégories phonétiques très voisines.

Pour les occlusives, on comprend tout l'intérêt d'associer à un système de reconnaissance un suivi des lèvres qui peut désambiguïser [p b] de [t k].

3.2. Cohortes CVCV

Leur fréquence représente un peu plus de 7% (3 115 occurrences) du total des cohortes. Si on distingue les représentants de la cohorte CVCV selon les classes définies par les macro-traits, on obtient 1 507 classes distinctes, soit 48% des cohortes. La distribution rang-occurrences (Figure 3) est quasi-linéaire sur une échelle Log (Log).

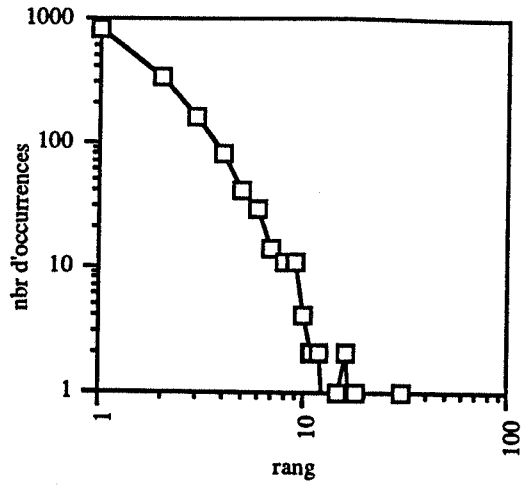


Figure 3 : Distribution rang-occurrences des cohortes CVCV

On retrouve à peu de détails près (très faible apparition de [s ʃ]) les mêmes macro-classes consonantiques et vocaliques que pour les CVCVCV. Il faut noter ici la très forte structuration qu'exerce la syllabe initiale CV sur le système sonore.

3.4. Cohortes CCVCV

Leur fréquence représente un peu plus de 3% (1 276 occurrences) du total des cohortes. Si on distingue les représentants de la cohorte CCVCV selon les classes définies par les macro-traits, on obtient 825 classes distinctes, soit 65% des cohortes.

Les groupes consonantiques en position initiale ont déjà été étudiés car ils permettent de dégager des régularités phonétiques [Rossi, 1968 ; Aubergé, 1991] (Liste 8).

10 [p t k] [j w l m n] [a ɛ œ̃ œ] [j w l m n] [e y]
8 [p t k] [j w l m n] [a ɛ œ̃ œ] [p t k] [e y]
8 [p t k] [ʀ v] [a ɛ œ̃ œ] [j w l m n] [e y]
6 [b d g] [ʀ v] [ɔ] [j w l m n] [e y]
6 [p t k] [ʀ v] [ɛ] [p t k] [e y]
6 [p t k] [ʀ v] [ɛ] [j w l m n] [ə]
6 [b d g] [j w l m n] [a ɛ œ̃ œ] [j w l m n] [ə]
6 [b d g] [ʀ v] [a ɛ œ̃ œ] [p t k] [e y]
6 [p t k] [j w l m n] [a ɛ œ̃ œ] [j w l m n] [ə]
6 [b d g] [ʀ v] [i] [j w l m n] [e y]

Liste 8 : Les occurrences les plus fréquentes des cohortes CCVCV de macro-traits identiques et les sons possibles

De ces comptages lexicaux émergent en tête [p t k b d g f] avec [ʀ] et [p f k b g] avec [l]. C'est-à-dire très majoritairement des occlusives (sourdes) avec les liquides [ʀ l]. On retrouve bien cette émergence avec les macro-classes les plus fréquentes.

3.5. Cohortes VCVCV

Leur fréquence représente un peu plus de 2% (1 113 occurrences) du total des cohortes. Si on distingue les représentants de la cohorte CCVCV selon les classes définies par les macro-traits, on obtient 820 classes distinctes, soit 74% des cohortes.

Pour les voyelles en position initiale, on retrouve deux catégories : la macro-classe [e y] c'est-à-dire les voyelles fermées mais non extrêmes (comme [i]) et des voyelles ouvertes, deux catégories très séparables (Liste 9).

8 [e y] [p t k] [e y] [ʀ v] [e y]
6 [e y] [p t k] [a ɛ œ̃ œ] [j w l m n ɲ] [e y]
5 [a ɛ œ̃ œ] [j w l m n ɲ] [e y] [j w l m n ɲ] [e y]
5 [a ɛ œ̃ œ] [j w l m n ɲ] [ɛ] [j w l m n ɲ] [ə]
5 [a ɛ œ̃ œ] [j w l m n ɲ] [ə] [j w l m n ɲ] [e y]
5 [a ɛ œ̃ œ] [s ʃ] [e y] [ʀ v] [e y]
5 [e y] [p t k] [a ɛ œ̃ œ] [j w l m n ɲ] [ə]
5 [ɑ̃] [p t k] [a ɛ œ̃ œ] [j w l m n ɲ] [e y]

Liste 9 : Les occurrences les plus fréquentes des cohortes VCVCV de macro-traits identiques et les sons possibles

4. CONCLUSION ET PERSPECTIVES

Cette première étude confirme un certain nombre de tendances universelles au niveau de la structuration de la séquence de sons : la forte domination des syllabes CV qui organise en position initiale la prédominance des occlusives sourdes [p t k] et dans une moindre mesure des sonores. Ces tendances sont nettement confirmées avec les groupes consonantiques. Ce sont les voyelles ouvertes qui constituent le noyau vocalique de la syllabe initiale, ce qui donne des syllabes très contrastées en termes d'énergie. Est mis en évidence, pour une même classe de cohortes et une même séquence de macro-traits, le petit nombre d'entrées lexicales ambiguës (inférieur à la dizaine dans la plupart des cas). Ces résultats semblent encourageants pour le filtrage lexical dans le cadre de la reconnaissance.

Le lexique des formes canoniques a permis de générer, en collaboration avec l'équipe TRILAN (LGI, Grenoble), le lexique orthographique des formes fléchies, soit environ 450 000 formes. Les formes phonétiques correspondantes vont être dérivées à partir de la grammaire de phonétisation obtenue sur lexique des formes canoniques [Belrhali et al., 1994]. Nous étendrons donc ensuite les fonctions macro-traits, cohortes, voisinages et points d'unicité à ce nouvel ensemble de données.

La taille du lexique des formes fléchies devient très difficile à gérer dans un environnement de type Macintosh. C'est pourquoi, nous avons établi un cahier des charges qui prévoit l'implantation du nouveau lexique dans un environnement Unix/ORACLE par le Laboratoire d'Informatique d'Avignon dans le cadre du GT lexique. La gestion de ce lexique devra répondre à deux contraintes essentielles (1) un accès rapide à chacun des champs, en conservant bien entendu le lien avec les informations associées (2) la possibilité de construire des sous-lexiques selon la demande de l'utilisateur. Il s'agit donc à présent de décider quels compromis entre les listes d'items et les fonctions de générations seront adoptés dans le lexique.

RÉFÉRENCES

Ait-Messaoud H. (1993) Catégorisation subphonétique et lexique du français. DEA Sciences du Langage, Université Stendhal, Grenoble.

Aubergé V. (1991) La synthèse de la parole : « des règles aux lexiques ». Thèse Informatique, Université P. Mendès France, Grenoble.

Aubergé V., Autesserre D., Bachut D., Béchet F., Belrhali R., Boë L.J., Courtin J., Dujardin D., Gerber R., Kandel S., Maret D., Méloni H., Plénat M. (1993) ORALL: Organisation, Representation and Access to a Large Lexicon. SALT Workshop "The Unified Lexicon", Durham.

Aubergé V., Boë L.J. & Lefèvre J.P. (1988) Lexique et groupe consonantiques. 17e JEP, GCP, 55-60.

Belrhali R., Aubergé V. & Boë L.J. (1994) La phonétisation du français : logique et linguistique dans une grammaire-lexiques. 20e JEP, Lannion.

Bernathova N. & Zackova P (1992) Typologie des occlusives et constrictives d'UPSID. DEA Sciences du langage, Université Stendhal, Grenoble.

Boë L.J. & Tubach J.P. (1992a) BDPHO : une base de données lexicale orthographique-phonétique du français parlé. Séminaire Lexique, IRIT-UPS Toulouse.

Boë L.J. & Tubach J.P. (1992b) "De A à Zut" Dictionnaire phonétique du français parlé. Grenoble, ELLUG.

Brodda B. (1979) Nagot om de Svenska Ordens Fonotax och Morfotax : Iakttagelser med Utgångspunkt från Experiment med Automatisk Morfologisk Analys. Papers from the Institute of Linguistics, Univ. Stockhom, 38, 63-108.

Bulot R. (1987) Techniques d'intelligence artificielle pour la reconnaissance de la parole : application au décodage acoustico-phonétique. Thèse d'Univ. d'Aix-Marseille II, Faculté de Luminy, Informatique.

Caelen J. (1979) Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique. Thèse de Doctorat d'état, Univ. Paul Sabatier, Toulouse.

Carlson R., Elenius K., Granström B. & Hunnicutt S. (1985) Phonetic and orthographic properties of the basic of five european languages. STL QPSR 1, 63-94.

Content A., Mousty P. & Radeau M. (1990) BRULEX une base de données lexicales informatisée pour le français écrit et parlé. L'Année Psychologique 90, 551-566.

De Mori R. (1983) Extraction of acoustic cues using a grammar of frames. Speech Com., 2, 2-3, 223-225

Dujardin D., Belrhali R., Boë L.J. & Courtin J. (1994) Lemmatisation du dictionnaire phonétique du français parlée « De A à Zut ». 20e JEP, Lannion.

Gilles P. (1993) Décodage phonétique de la parole et adaptation au locuteur. Thèse Informatique, Université d'Avignon et des pays de Vaucluse.

Halle M. & Vergnaud J.R. (1980) Three-dimensional phonology. Jounal of Linguistic Research.

Jakobson R., Fant G. & Halle M. (1952) Preliminaries to speech analysis, the distinctive features and their correlates. MIT, Cambridge, Technical report 13.

Kolinsky R., Morais J. & Segui J. (1991) La reconnaissance des mots dans les différentes modalités sensorielles : études de psycholinguistique cognitive. PUF, Paris.

Lindblom B.E.F. (1983) Economy in speech gesture. In *The Production of Speech*, P.F. Mac Neilage ed., 217-245, Springer-Verlag, New-York.

Maddieson I. (1993) The structure of segment sequences. UCLA Working papers in phonetics, 83, 1-8

Maddieson I. & Precoda K. (1992) Phonetics models and syllable structure. Phonlogy 9, 45-60.

Marslen-Wilson W.D. (1984) Function and process in spoken word recognition. In H. Bouma & D.G. Bouwhuis (Eds), Attention and Performance X : Control of Language Processes, Hillsdale, NJ, Lawrence Erlbaum Associates, 125-149.

Marslen-Wilson W.D. (1987) Function parallelism in spoken word-recognition. Cognition, 25, 1-2, 71-102.

Méloni H. (1982) Étude et réalisation d'un système de reconnaissance automatique de la parole. Thèse de Doctorat d'état, Faculté de Luminy, Marseille.

Mercier G. (1983) Le décodage acoustico-phonétique de la parole. Bulletin de l'INRIA, 84.

Pérennou G. & de Calmès M. (1986) BDLEX. : Une base de données et de connaissances du français parlé. Séminaire Lexique, Toulouse, 243-258.

Pérennou G. (1988) Le projet BDLEX de base de données et de connaissances lexicales et phonologiques. GRECO-PRC, Paris, 81-111.

Rossi M. (1968) Au sujet des groupes consonantiques du français. Revue d'acoustique, 3-4, 306-311.

Rossi M., Nishinuma Y. & Mercier G. (1983) Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnassance automatique de la parole. Speech Communication, North-Holland, 2, 215-217.

Togeby K. (1965) Structure immanente de la langue française. « Langue et Langage » Larousse, Paris.

Wiren J. & Tubbs H.L. (1956) Electronic binary selection system of phonem classification. J. Acoust. Soc. Amer. 28, 1082-1091.

Zipf G.K. (1935) The psycho-biology of language. Hough, Houston.

UTILISATION CONJOINTE DE CONNAISSANCES ET DE TECHNIQUES NEUROMIMÉTIQUES POUR LA RECONNAISSANCE DU LOCUTEUR

QUÉLAVOINE REGIS, NOCERA PASCAL ET BONASTRE JEAN FRANÇOIS.

LABORATOIRE D'INFORMATIQUE DE L'UNIVERSITE D'AVIGNON ET DES PAYS DE VAUCLUSE.

Résumé

Nous proposons ici une méthode utilisant des réseaux de neurones multicouches pour la reconnaissance du locuteur. L'architecture adoptée permet de prendre en compte l'ajout de nouveaux locuteurs de façon dynamique, sans avoir à reprendre les apprentissages déjà effectués, à travers la construction d'un arbre de décision dans lequel chaque noeud est un réseau de neurones, et chaque feuille un locuteur. Une branche de l'arbre symbolise une classe de locuteurs ayant des voix proches. En outre, nous localisons de façon automatique les zones intéressantes du signal avant de les présenter aux réseaux, ce qui nous permet d'obtenir des taux de reconnaissance de 80% en moyenne, sur de courts enregistrements (1 à 2" de parole) et avec 47 locuteurs.

Mots clés: reconnaissance du locuteur, réseaux de neurones, informations explicites.

1. INTRODUCTION

Le thème abordé dans cet article est l'identification, au sens informatique, d'un locuteur; cette tache consiste à déterminer l'identité d'une personne coopérative caractérisée par un échantillons de sa voix, sachant qu'elle appartient à un ensemble de locuteurs connus du système.

Si nous sommes habituellement capables de reconnaître les voix d'un assez grand nombre de familiers, nous sommes par contre peu habiles à décrire quelles informations nous le permettent: timbre, intonation, rythme, ou encore choix des mots, etc... Différentes approches ont été etudiées, en s'intéressant à des portions plus ou moins longues de parole, afin de calculer des spectres moyens, d'étudier les variations de la F0, ...

Parmi les différentes techniques utilisées, les méthodes neuromimétiques et algébriques sont très performantes [Bennani, 92; Bimbot, 93; Montacié, 93]. La plupart nécessitent toutefois des messages relativement longs (de 2 à 10") ou encore des apprentissages importants, que ce soit en nombre d'échantillons ou en ressources machines. De plus, l'ajout d'un nouveau locuteur impose souvent de reprendre la totalité de l'apprentissage (en particulier dans le cas des méthodes connexionistes). D'autres approches ont montré que l'utilisation de connaissances explicites pour sélectionner et traiter l'information pertinente contenue dans le signal [Bonastre, 93, 94], permettait d'accroitre les performances sans présenter ces inconvénients.

Cependant, lorsque nous savons localiser à l'aide de connaissances expertes l'information pertinente, il est souvent difficile de la traiter explicitement de manière optimale. Par contre, les réseaux de neurones, à travers l'apprentissage par l'exemple, offrent une solution intéressante à ce problème. Utilisés conjointement avec les connaissances de l'expert, ils peuvent compenser les lacunes de notre savoir [Nocera, 90, 92]. Pour faciliter l'ajout d'un locuteur (sans remettre en cause le travail déjà effectué), Artières [Artières, 93] propose d'utiliser des réseaux modélisant chacun les caractéristiques d'une personne; les performances sont intéressantes sur une base restreinte mais restent à démontrer pour un grand nombre de locuteurs. En effet, l'erreur moyenne de prédiction entre deux personnes risque d'évoluer proportionnellement au nombre de locuteurs. De plus, la reconnaissance requiert l'utilisation de tous les réseaux. Une alternative consiste à construire un ensemble de réseaux séparant chacun deux personnes. Rudasi [Rudasi, 93] crée toutes les combinaisons possibles (soit pour n individus, n*(n-1)/2 réseaux), ce qui interdit l'application de cette technique dès lors que l'on atteint un nombre conséquent de locuteurs. Nous proposons de reprendre ce principe en représentant l'ensemble des locuteurs par un nombre réduit de réseaux structurés sous la forme d'un arbre binaire de décision.

2. LES DONNÉES

La base des locuteurs est constituée par une cinquantaine de personnes coopérantes mais novices à cet exercice, enregistrées dans un milieu bruité de type bureau. Chaque locuteur a lu des phrases issues de la base BDSONS: 50 sont extraites du corpus PEQ pour l'apprentissage et 44 du corpus NAS pour la reconnaissance.

Les enregistrements sont alors traités par L.P.C. (14 coefficients) de façon à retenir, tous les centièmes de seconde, un spectre sur 24 canaux répartis suivant une échelle de Mel. Une segmentation automatique de l'énoncé est ensuite effectuée à partir de la transcription phonétique normative de l'énoncé, permettant d'obtenir la chaîne phonémique prononcée la plus probable [Bonastre, 94]. Pour chaque locuteur, nous retenons les portions de signal contenant un phonème décrit par trois spectres (un pour le phonème central et un pour chaque transition [Gilles, 93; Méloni, 91]). La variété des phrases utilisées ne permet de couvrir qu'environ 10% des triplets phonétiques du français. La segmentation automatique, combinée avec des erreurs de lecture pour certains locuteurs, diminue la qualité des corpus.

3. LES RÉSEAUX MULTICOUCHES

Pour un problème de classification en mode supervisé, les réseaux multicouches utilisés avec la retro-propagation du gradient (ou ses variantes), sont parmi les solutions neuromimétiques les mieux adaptées [Caudill, 91]. De plus, le potentiel de généralisation de ces derniers est d'autant plus élevé que le nombre de classes à séparer est faible [Bochereau]. Ceci nous a conduit à construire un arbre de séparateurs dichotomiques, très efficace sur un nombre restreint de locuteurs [Quélavoine, 92].

Les réseaux utilisés sont à trois couches :

- 72 noeuds en entrée (3 vecteurs de 24 paramètres correspondant au phonème étudié et à ses contextes gauche et droit),
- 20 noeuds en couche cachée,
- 2 noeuds en sortie (locuteur α ou β).

Chaque couche est complètement connectée à la suivante.

3.1 Reconnaissance

En situation de reconnaissance, chaque unité phonétique de la phrase testée est présentée en entrée du réseau constituant la racine de l'arbre de décision. De façon récursive, elle est alors orientée vers le sous arbre droit ou gauche suivant la meilleure réponse du réseau, jusqu'à arriver au niveau des feuilles, représentant chacune un locuteur.

Sur l'ensemble de la phrase, nous comptons alors, pour chaque locuteur, combien d'unités lui ont été attribuées. La personne reconnue est celle ayant le meilleur score.

3.2 L'arbre

La construction de l'arbre se fait au fur et à mesure de l'ajout de nouveaux locuteurs.

Ainsi, le premier noeud de l'arbre est un réseau séparant le premier locuteur du deuxième.

Le principe d'évolution est le suivant: étant donné l'arbre à l'étape n-1 (il y a déjà n locuteurs traités), nous passons le corpus d'apprentissage du locuteur n+1 en reconnaissance. Cette personne étant inconnue à ce moment là, les réponses de l'arbre permettent de connaître les locuteurs confondus avec le nouvel arrivant.

Nous sélectionnons alors les personnes i par ordre décroissant des pourcentages de confusion, jusqu'à obtenir 70% des cas, et nous remplaçons les feuilles correspondantes par des réseaux séparant chaque locuteur i avec l'individu n+1.
Nous avons dans la figure 1 l'exemple où c peut être alternativement confondu avec a ou b.

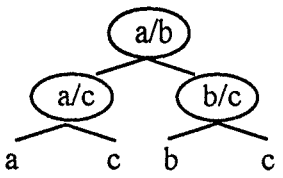


fig.1: exemple de structure d'arbre

La figure 2 montre l'évolution de l'arbre lors du rajout du locuteur d (confondu uniquement avec a). L'arbre ainsi construit a une profondeur maximale égale au nombre de locuteurs.

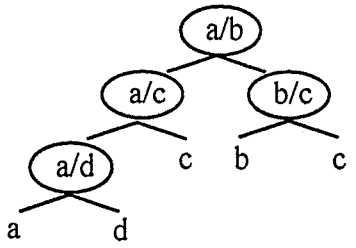


fig.2: ajout d'un locuteur

Le nombre maximum de réseaux est déterminé par le pourcentage cumulé de confusion pris en compte lors de l'apprentissage (70% dans notre cas). Nous arrivons alors à une moyenne de 5 réseaux par locuteur (3 au mieux, 10 au pire). En effet, nous observons que les locuteurs se regroupent plutôt par classes qui forment des branches distinctes de l'arbre.

De manière à restreindre le nombre de réseaux, nous pourrions ne plus limiter la croissance de l'arbre au niveau des seules feuilles, mais l'autoriser à tous les niveaux. Si un locuteur c peut être confondu avec a et b, nous construirons alors un séparateur ab/c. La figure 3 reprend le graphe de la figure 2 avec ce principe. Toutefois, cela alourdit les réseaux, qui auront des corpus d'apprentissage plus importants et déséquilibrés.

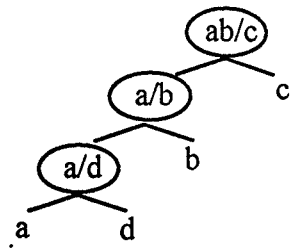


fig.3: exemple modifié de structure d'arbre

3.3 Apprentissage

Nous réalisons l'apprentissage de chacun des réseaux de l'arbre de façon indépendante. Le corpus d'apprentissage d'un réseau est uniquement constitué à partir des exemples correspondant aux deux locuteurs à séparer. Sa taille reste raisonnable et la représentation des classes est équilibrée (nous ne risquons pas le phénomène d'oubli comme dans les réseaux de type un contre tous).

Afin de prendre en compte le recouvrement des espaces de variation intra- et inter-locuteurs, nous considérons l'apprentissage comme terminé lorsque pendant dix itérations, plus de 80% des trames de chaque locuteur du corpus d'apprentissage ont bien été affectées à leur propriétaire.

4. RÉSULTATS

En ne se servant que d'une seule phrase pour reconnaître une personne (soit 1 à 2"s de parole effective), nous obtenons une moyenne de 80% de réussite en première position, 90% sur les deux premiers locuteurs proposés, avec une base de 47 personnes. Ceci est très prometteur si nous nous rappelons que seulement 70% des confusions sont prises en compte pour la construction de l'arbre.

5. CONCLUSIONS

La technique proposée permet de rajouter de façon aisée un nouveau locuteur, sans remettre en cause les apprentissages antérieurs. Nous obtenons de bons scores en reconnaissance à partir de messages de courte durée (80% pour 2" et 47 locuteurs). Ces résultats confirment que les réseaux de neurones sont de bons classifieurs pour la reconnaissance du locuteur, dont les capacités sont d'autant mieux exploitées que l'information qu'ils traitent est pertinente.

Une meilleure construction de l'arbre pour limiter le nombre de réseaux utilisés, mais aussi une optimisation de chacun de ces derniers, devrait nous permettre d'accroître les taux de reconnaissance.

En effet, pour les personnes ayant les scores les moins élevés, nous pouvons aisément localiser les erreurs et étoffer en conséquence le corpus d'apprentissage sans avoir à le reprendre pour tous les locuteurs. Effectivement, les erreurs sont souvent dues à un manque d'exemples de phonèmes dans certains contextes, suite à une mauvaise segmentation des phrases d'un locuteur.

L'information que nous utilisons pour caractériser un individu reste statique (parties stables des phonèmes) malgré une prise en compte des transitions inter-phonémiques. Or, une partie importante des spécificités individuelles véhiculées par la voix est de nature dynamique. La recherche d'un système d'étude exploitant ces informations devraient améliorer encore les résultats.

Toutefois, si notre système s'avère efficace sur une cinquantaine de locuteurs, il convient de le tester sur une base plus importante de manière à vérifier que le nombre de nouveaux réseaux nécessaires pour l'ajout d'une personne n'augmente pas proportionnellement au nombre de locuteurs, tout en conservant les mêmes performances.

6. BIBLIOGRAPHIE

Artières T. et Gallinari P.: *Neural models for extracting speaker characteristics in speech modelization systems.* 3rd Eurospeech, Berlin, Germany, pp.2263-2266, 1993.

Bennani Y.: *Approches connexionistes pour la reconnaissance automatique du locuteur: modélisation et identification.* Thèse de l'Université Paris Sud, 1992.

Bimbot F. et Mathan L.: *Text-Free speaker recognition using an arithmetic-harmonic sphericity measure.* 3rd Eurospeech, pp.169-172, Berlin, Germany, 1993.

Bochereau L. et Bourgine P.: *A generalist-specialist paradigm for multilayer neural networks.* CEMAGREF, Antony.

Bonastre J.F.: *Stratégie analytique orientée connaissances pour la caractérisation et l'identification du locuteur.* Thèse de l'université d'Avignon et des Pays de Vaucluse, jan94.

Bonastre J.F et Méloni H.: *Speaker recognition and analytic process.* 3rd Eurospeech, Berlin, Germany, 1993.

Caudill M.: *Avoiding the great backpropagation trap.* AI expert, vol.6, n°7,jul91, pp. 29, 30, 32, 34, 35.

Gilles P.: *Décodage phonétique de la parole et adaptation au locuteur*. Thèse de l'université d'Avignon et des Pays de Vaucluse, jan93.

Méloni H. et Gilles. P.: *Décodage acoustico-phonétique ascendant*. Traitement du Signal, vol.8, n°2, pp.107-114, 1991.

Montacié C. et Le Floch J.L.: *Discriminant AR-vector models for free-text speaker verification*. Proc. 3rd Eurospeech , Berlin, Germany, pp.161-164, 1993.

Nocera P.: *Utilisation conjointe de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique*. Thèse de l'université d'Avignon et des Pays de Vaucluse, jan92.

Nocera P.et Bulot R.: *Rule driven neural networks for speech recognition*. INNC90, Paris, vol.1, pp.191-194, 1990

Quélavoine R.: *Identification des locuteurs à l'aide de réseaux neuromimétiques*. Mémoire de DEA de l'université d'Avignon et des Pays de Vaucluse, sep92.

Rudasi L. et Zahorian S.A.: *Text independent speaker identification using binary-pair partitioned neural networks*. ICASSP 93, vol.4, pp679-684 .

LES PHÉNOMÈNES DE VARIATION DE GESTE : ÉTUDE DES VARIATIONS ARTICULATOIRES DE GROUPES CONSONANTIQUES SOUS L'EFFET DU DÉBIT, EN FRANÇAIS

Béatrice VAXELAIRE

Institut de Phonétique-Université de Strasbourg II
22 rue Descartes-67084 Strasbourg Cédex-France

ABSTRACT

The aim of this paper is the study of the effects of speech rates -slow, fast- on the articulatory parameters. This work is to evaluate with the technique of cineradiography the comparative articulatory behavior of consonant clusters in unstressed position for 14 parameters. This examination shows that fast rate implies a few modification of the articulatory gestures. There is an adaptation of the articulators to the speed of the movement through new movements, aggregation and new anticipatory gestures.

1. INTRODUCTION

La direction principale de notre travail concerne l'étude des phénomènes de variation du geste articulatoire sous l'effet du débit. Il s'agit d'une description quantitative de la variabilité spatio-temporelle des mouvements articulatoires de la parole, avec une variation contrôlée de la vitesse d'élocution. L'organisation spatio-temporelle dans les processus de la production de la parole est un problème central pour l'étude de la communication parlée et de la phonétique. Notre étude fournit des données sur la variabilité articulatoire associée aux variations de débit en français, et pourrait contribuer à la compréhension du contrôle de la parole lié au "paramètre débit".

La variation de l'organisation se manifeste sur des positions d'articulateurs tels que la langue, les lèvres ou le voile du palais. Les mouvements coordonnés constituent le geste et représentent une déformation temporaire du conduit vocal [14]. Ainsi, les gestes sont utilisés à la fois pour caractériser des données de mouvements articulatoires et pour constituer une base d'unités phonologiques. Les gestes sont capables de réaliser des adaptations nécessaires aux contraintes liées aux mécanismes productifs et perceptuels des phénomènes de communication. Le niveau articulatoire nécessite une observation rigoureuse du conduit vocal et de ses comportements parfois très complexes. Le conduit vocal est constitué de différentes parties individuelles coordonnées en un tout unifié, requérant une orchestration de ses différentes parties, dans le but d'atteindre les cibles. Les travaux en articulatoire sur l'observation d'une variation de débit sont finalement rares, ou bien ils concernent une zone partielle et isolée du niveau articulatoire et ne sont pas représentatif de sa globalité, ou bien ils concernent d'autres langues que le français [7, 8, 9, 10 et 23]. La motivation première de notre présente investigation est, compte tenu de la carence en données articulatoires sur les effets du débit, de recueillir en un premier temps des renseignements quant aux phénomènes comportementaux des articulateurs et à leur variations d'un débit à l'autre. Le but que nous nous sommes fixé au départ de cette étude est une description phonétique et phonologique d'une variation de débit. La chaîne parlée observée à travers une perturbation de débit montre au-delà d'une réduction de la durée, une modification du schéma dynamique des articulateurs qui adoptent une nouvelle "stratégie" en rapport avec le débit [24]. Sous l'effet du débit, le programme moteur est préparé à l'avance pour permettre aux articulateurs l'exécution de chacune des unités, et nous observons une variation de l'acte de parole [21]. La chaîne parlée n'est pas touchée de la même manière par le débit : la compression de la durée des phonèmes est non-uniforme entre les consonnes et les voyelles, ce qui est une manifestation de la réorganisation articulatoire due au changement de débit. Il existe entre les articulateurs une "interaction non-linéaire" pouvant s'interpréter par une sorte d'"autorégulation interne" [3]. Il existe une compensation articulatoire entre les différents articulateurs dans la parole naturelle, et on ne peut parler de "programmation motrice", et donc de production de la parole, sans prendre en compte le phénomène de "coordination" entre les articulateurs. Les commandes motrices sont en relation avec les variabilités articulatoires et acoustiques [12 et 13]. Les réflexions et les études sur les phénomènes de coarticulation font

constamment référence à la position des organes de la parole et aux tendances motrices responsables des habitudes articulatoires et phoniques. C'est la "base articulatoire" [22].

Par leur nature, les voyelles imposent un contrôle articulatoire sur les positions linguales pour leur ensemble [2], alors que les consonnes n'influencent qu'une seule partie de la langue, à savoir celle concernée par le lieu de la constriction ou de l'occlusion [17]. Le degré de fusion de deux éléments occlusifs linguaux, lors de chevauchements entre différentes régions linguales, témoignent de la contrainte articulatoire au sein du groupe de consonnes, expliquant l'arrêt des effets coarticulatoires de la voyelle sur ce groupe de consonnes [18].

L'observation de la parole sous deux débits différents montre de nombreuses modifications tant sur les plans articulatoire et acoustique que sur le plan perceptif. Un débit rapide entraîne des phénomènes de réduction : au niveau articulatoire il est observé une modification des mouvements dans leur durée et leur amplitude, et au niveau acoustique il se produit une réduction de la durée de la phrase, de la syllabe et du phonème ainsi qu'une réduction formantique des voyelles. L'effet de réduction est dû à la contiguïté voire au chevauchement spatio-temporels des phonèmes adjacents sous l'action des commandes motrices [11]. En débit rapide, la dynamique des articulateurs est modifiée et peut entraîner des modifications de phase, ainsi qu'une optimisation dans la coordination des gestes entre les mouvements vocaliques et consonantiques [6]. Il peut se produire un phénomène de neutralisation de la quantité [5].

2. MÉTHODE

2.1 Présentation

Notre but étant de décrire les manifestations articulatoires de la répercussion d'une variation de débit dans la chaîne parlée, nous avons choisi la **cinéradiographie**. Cette technique allie l'analyse du développement des mouvements articulatoires en montrant l'ensemble du travail des articulateurs par le déroulement du film à la vitesse de 50 images/seconde, saisissant ainsi les déplacements des organes toutes les 20 ms, avec son synchrone, à l'étude des positions statiques par l'utilisation de l'arrêt sur image. Les vues cinéradiographiques de profil nous montrent la succession temporelle des contours du conduit vocal dans le plan sagittal médian. La cinéradiographie est utilisée à l'*Institut de Phonétique Générale et Expérimentale de Strasbourg*, et a permis la description articulatoire des systèmes phonétiques [20 et 4], l'observation des phénomènes de coarticulation [1], des traits de labialité [26] ou encore de la position cible [19].

2.2 Corpus

Notre corpus a été élaboré pour une étude originelle plus générale sur l'observation comparée entre les débits lent et rapide des consonnes occlusives orales, sourdes et sonores, du français. Le corpus complet représente 58 phrases courtes, de 4 à 6 syllabes, chargées de sens, enregistrées en débits lent et rapide par deux sujets français (un homme et une femme), dont la parole pourrait ressembler au français de la "norme". Le corpus propose un certain nombre de séquences du type VCV consonne simple, ou du type VCCV consonne double ou groupe de deux consonnes différentes, tous deux avec un entourage vocalique identique et unique ([aCa] ou [aCCa]), ou un entourage labialisé ou nasal (l'élément labialisé ou nasal est placé avant ou après la consonne), ou en groupes de consonnes courants en français [25].

Pour cette présente étude concernant plus particulièrement les variations articulatoires de groupes consonantiques sous l'effet du débit, et plus précisément le phénomène d'agrégation de gestes, nous avons choisi de décrire 4 segments différents en débit lent et en débit rapide, correspondant au sujet masculin. Nous avons retenu 3 phrases du corpus (les groupes de consonnes observés sont soulignés).

1 Les droits octroyés [ledʀwazɔktʀwáje]
2 La patte large broyée [lapatlaʀʒbʀwáje]
3 Le petit âne cassé [ləptitankáse]

2.3 Paramètres articulatoires

Nous avons choisi une image de référence dans le radiofilm à partir de laquelle nous avons établi notre grille de mesures dans un repère orthonormé, avec les points fixes des dents de la mâchoire supérieure ainsi que ceux du palais dur.

Les paramètres articulatoires sont au nombre de 14 (fig. 1) :

1. projection de la lèvre supérieure (LS)
2. projection de la lèvre inférieure (LI)
3. écartement bilabial (EB)
4. angle des maxillaires (AM)
5. 6. 7. 8. 9. élévation de la langue, de l'avant vers l'arrière
10. racine de la langue

11. voile du palais : hauteur maximale, écartement de la paroi pharyngale et étendue de contact voile/ paroi pharyngale
12. os hyoïde : mouvement horizontal et vertical (OH)
13. base du larynx (BL)
14. épiglotte : mouvement horizontal et vertical (ÉP)

Nous avons également relevé l'étendue de contact pendant l'occlusion de la langue ou des lèvres, correspondant au lieu d'articulation.

2.4 Débits

Le corpus a été enregistré en parole lue, sans pause silencieuse à l'intérieur des phrases ni marquage intonatif particulier. Nous pouvons parler de suites sonores et de débit d'élocution pour vitesse de parole, ainsi que de temps d'articulation pour temps de locution. Les ruptures de vitesse ou de rythme ont été évitées, et les modalités de l'enregistrement ont contribué à la régularité du débit. Nous devons préciser que lors de l'enregistrement en salle de radiologie, nous sommes contraints de faire un arrêt toutes les vingt secondes.

Le débit lent (DL) correspond à une vitesse de lecture qui se rapproche de la vitesse de dictée d'un message. Le débit rapide (DR) correspond à une parole accélérée, dont la vitesse permet, malgré la rapidité d'émission, une bonne compréhension par l'interlocuteur, c'est-à-dire que le message est perçu sans que les sons soient réalisés pour autant. En effet, l'interlocuteur est capable de percevoir des sons qui n'ont pas étés émis par le locuteur, grâce à ses possibilités cognitives et à cause de ses habitudes linguistiques [15].

Pour les 3 phrases présentées ici, nous obtenons en moyenne les vitesses de parole suivantes :
DL : 4,5 syllabes/seconde
DR : 6,4 syllabes/seconde (fig. 2)

3. VARIATION ARTICULATOIRE

3.1 Paramètres 1, 2, 3 et 4

Les mouvements des lèvres supérieure et inférieure, projection et écartement, ainsi que le mouvement des maxillaires montrent une différence de comportement entre les débits lent et rapide :
- la protrusion labiale est inférieure (de 2 à 4 mm) en DR et en contexte non-labialisé, ce qui justifierait une tension musculaire supérieure pour DR.
- L'aperture de la voyelle impose au groupe consonantique en DR, une contrainte à l'écart interlabial et à l'angle des maxillaires : EB est plus petit (de 6 mm) en DR pour [t]+[i] (ph. 3), et il est plus grand pendant les occlusions intra-buccales des consonnes ayant un entourage en [a] ;

AM est influencé par l'aperture des voyelles constituant l'entourage consonantique de manière plus importante en DR qu'en DL.
- L'étendue du contact occlusif bilabial est inférieure en DR pour [p] (ph. 3), ce qui est contraire au principe de tension relevé par ailleurs (infra : étendue de contact occlusif pour les intra-buccales), ce qui pourrait s'expliquer par un contact plus rapide et donc moins appuyé.

3.2 Paramètres 5 à 10 et 11

Les paramètres 5 à 10 représentent les points de la langue, de l'apex à la racine ; le paramètre 11 correspond au voile du palais, hauteur, recul et étendue de contact avec la paroi pharyngale.
- L'aperture de la voyelle impose au groupe consonantique en DR, une contrainte à l'élévation de la langue : en effet, malgré une grande variabilité, la langue a une tendance renforcée en débit rapide à un positionnement en fonction de l'entourage vocalique.

Ainsi, la langue est plus élevée en DR après l'explosion de la vélaire et pendant l'occlusion alvéodentale pour le groupe [kt] (ph. 1) comme pour le groupe [pt] (ph. 3) ; et inversement, la langue est moins élevée en DR pour le groupe [tl] (ph. 2).
- La racine de la langue, dans son mouvement horizontal, a tendance à avoir une même réaction anticipatrice que la masse linguale.

Ainsi la racine, tout en ayant une amplitude de mouvement réduite en DR, aura une position plus reculée si le son subséquent est postérieur, par exemple [kt]+[ʀw] ; et inversement une position plus avancée avec une voyelle subséquente antérieure, par exemple [pt]+[i] pour lequel la racine s'avance également de 4 mm en DL et en DR, avec cependant une position de départ plus avancée de 4 mm en DR.
- L'étendue de contact occlusif, pour les occlusives intra-buccales alvéo-dentales et vélaires, est supérieure en DR, résultat de l'amortissement de la force articulatoire. L'impact de la langue est plus important parce que l'organe arrive plus vite. En effet, il semble que la raideur articulatoire, nécessitée par la rapidité du mouvement, a pour conséquence de propulser l'articulateur avec une plus grande force vers sa cible [16].
- les occlusions dans chaque groupe de consonnes sont successives en DL, avec 1, 2 ou 3 images, représentant 20, 40 ou 60 ms d'intervalle, entre les deux occlusions, et

elles deviennent partiellement simultanées en DR sur 1 à 2 images, c'est-à-dire pendant 20 à 40 ms, pour les groupes d'occlusives de lieux d'articulation différents intra- ou extra-buccal.
La durée des occlusions est de :
- pour [kt] 40+20 ms avec un intervalle de 60 ms entre les 2 occlusions en DL, et de 100 ms en DR décomposées en 80+40 ms avec un chevauchement de 20 ms ;
- pour [tl] 40+40 ms sans image d'intervalle mais avec une explosion entre les 2 sons en DL, et de 80 ms sans explosion en DR ;
- pour [pt] 40+40 ms avec un intervalle de 40 ms entre les 2 occlusions en DL, et de 80 ms en DR décomposées en 40+60 ms avec un chevauchement de 20 ms ;
- pour [nk] 80+40 ms sans intervalle entre les 2 sons, et de 80 ms décomposées en 60+60 ms avec un chevauchement de 40 ms.
Pour le groupe [tl] (ph. 3), l'occlusive alvéo-dentale suivie de la constrictive latérale présentent le même schéma que les groupes d'occlusives, à savoir deux occlusions sur la ligne médiane de la vue de profil successives et séparées par 1 image en DL, le tout avec une durée de 140 ms ; et une seule occlusion située dans la région alvéo-dentale en DR avec une durée de 80 ms.
- le voile du palais, sous l'effet du tension musculaire nécessitée par la vitesse de parole, est projeté vers le haut et vers l'arrière contre la paroi pharyngale. Ainsi, son étendue de contact avec celle-ci est supérieure en DR pour les non-nasales.
Pour le groupe [nk] (ph. 3), le voile se relève plus rapidement en DR.
(remarque : nous avons remarqué dans le corpus étendu une tendance générale à l'anticipation du mouvement du voile en DR, également avec un abaissement prématuré pendant une occlusive non-nasale suivie d'un son nasal).

3.3 Paramètres 12, 13 et 14

Les mouvements de l'os hyoïde, de la base du larynx et de l'épiglotte montrent des différences de comportement articulatoire entre DL et DR.
- OH est généralement plus élevé et légèrement plus reculé en DR, ce qui représente un indice du renforcement articulatoire.
- BL semble fonctionner avec OH en étant généralement plus élevée en DR.
- EP a un comportement dépendant du paramètre 10, et suit la racine de la langue. (fig. 3)

4. CONCLUSION

L'augmentation de la vitesse d'élocution perturbe la dynamique articulatoire dans la production de la parole en déclenchant de nouvelles stratégies consécutives au renforcement articulatoire. Un mouvement rapide provoque une réduction de la durée et nécessite une force articulatoire particulière sur le plan de la dynamique articulatoire comme sur celui de la programmation. La modification articulatoire crée des phénomènes d'anticipation liées aux contraintes coarticulatoires, mais aussi une agrégation de gestes. Ainsi, les occlusions se chevauchent et l'étendue du contact occlusif est supérieure en débit rapide.

RÉFÉRENCES

[1] BIERMANN FISCHER M., 1990, *Analyse articulatoire des consonnes constrictives du français. Contribution à l'étude de la coarticulation à la lumière de la cinéradiographie.* Thèse de Doctorat Nouveau Régime, Institut de Phonétique, Université de Strasbourg.
[2] BOË L.-J. & al., 1992, "The geometric vocal tract variables controlled for vowel production : proposal for constraining acoustic-to-articulatory inversion", *Journal of Phonetics, vol. 20,* 27-38.
[3] BOTHOREL A., 1983, "Contraintes physiologiques et indices articulatoires", *Speech Communication, vol. 2, N^OS 2-3,* 119-122.
[4] BOTHOREL A., SIMON P., WIOLAND F. & ZERLING J.-P., 1986, "Cinéradiographie des voyelles et des consonnes du français, recueil de documents synchronisés pour quatre sujets : vues latérales du conduit vocal, vues frontales de l'orifice labial données acoustiques", *Travaux de l'Institut de Phonétique de Strasbourg,* 296 p..
[5] BOUSSAFFA F., JOMAA M. & SOCK R., 1991, "Les contraintes temporelles des types consonantiques sur le timing mandibulaire de la quantité", *Actes du XII^e Congrès International des Sciences Phonétiques, 19-24 août, vol. 2, Aix-en-Provence, France,* 306-309.
[6] DELATTRE C. & PERRIER P., 1991, "Phase modifications in tongue movements across speech rate variations : influence of consonantal articulation", *Actes du XII^e Congrès International des Sciences Phonétiques, 19-24 août, vol. 3, Aix-en-Provence, France,* 26-29.
[7] GAY T., 1978, "Effect of speaking rate on vowel formant movements", *JASA, 63/1,* 223-230.
[8] GAY T., 1981, "Mechanisms in the control of speech rate", *Phonetica, 38,* 148-158.

[9] GOPAL H. S., 1990, "Effects of speaking rate on the behavior of tense and lax vowel durations", *Journal of Phonetics, 18,* 497-518.
[10] JOMAA M. & ABRY C., 1988, "La résistivité de la quantité vocalique aux variations de la vitesse d'élocution. Le cas de l'arabe tunisien", *XVII^e JEP, Nancy.*
[11] LINDBLOM B., 1963, "Spectrographic study of vowel reduction", *JASA, 35,* 1773-1781.
[12] MAEDA S., 1989, "Articulation compensatoire des voyelles : analyse de données cinéradiographiques avec un modèle linéaire" *Département de Recherche en Communication par la Parole, CNET, Lannion,* 540-558.
[13] MAEDA S., 1991, "On articulatory and acoustic variabilities : implications for speech motor control", *PERILUS, XIV,* 79-82.
[14] MATTINGLY I. G., 1990, "The global character of phonetic gestures" *Journal of Phonetics, 18,*445-452.
[15] MILLER J. L., 1981, "Effects of speaking rate on segmental distinctions", *Perspectives on the Study of Speech Ed. by P. D. Eimas and J. L. Miller,* 39-74.
[16] PERRIER P., ABRY C. & KELLER E., 1989, "Vers une modélisation du dos de la langue" *Journal d'Acoustique, 2,* 69-77.
[17] RECASENS D., 1989, "Long range coarticulatory effects for tongue dorsum contact in VCVCV sequences", *Haskins Lab. Status Report on Speech Research, SR 99/100,* 19-27.
[18] RECASENS D., FONTDEVILA J. & PALLARES M. D., 1991, "Articulatory control in stop consonant clusters", *PERILUS, XIV,* 39-46.
[19] SCHELLENBAUM-PASTOR A., 1992, *Contribution à l'étude des voyelles [i, e, a] du français : variabilité articulatoire, recherche de l'image centrale, rapports interarticulateurs à partir de la cinéradiographie,* Thèse de Doctorat Nouveau Régime, Institut de Phonétique, Université de Strasbourg, 450 p..
[20] SIMON P., 1967, *Les consonnes françaises, mouvements et positions à la lumière de la cinéradiographie,* Thèse d'État Klincksieck, Paris, 380 p..
[21] STERNBERG S., KNOLL R. L., MONSELL S. & WRIGHT C. E., 1988, "Motor programs and hierarchical organization in the control of rapid speech", *Phonetica, 45, n^os 2-4,* 175-197.
[22] STRAKA G., 1989, "Base articulatoire. Essai d'une mise au point", *in Mélanges de Phonétique Générale et Expérimentale offerts à Péla Simon, publications de l'Institut de Phonétique de Strasbourg,* 757-768.
[23] STRANGERT E., 1987, "Speech rate and the temporal structure of stressed and unstressed syllables in Swedish", *Ruul, 17, Swedish phonetics conference,* 22-28.
[24] VAXELAIRE B., 1991, "De l'analyse d'une variation de débit dans la chaîne parlée, à la lumière de la cinéradiographie", *Actes du XII^e Congrès International des Sciences Phonétiques, 19-24 août, vol. 4, Aix-en-Provence, France,* 26-29.
[25] VAXELAIRE B., 1993, *Étude comparée des effets des variations de débit -lent, rapide- sur les paramètres articulatoires à partir de la cinéradiographie (sujets français),* Thèse de Doctorat Nouveau Régime, Institut de Phonétique, Université de Strasbourg, 530 p..
[26] ZERLING J.-P., 1990, *Aspects articulatoires de la labialité vocalique en français. Contribution à la modélisation à partir de labiographes, labiofilms et films radiologiques. Étude statistique, dynamique et contrastive,* Thèse d'État, Université de Strasbourg.

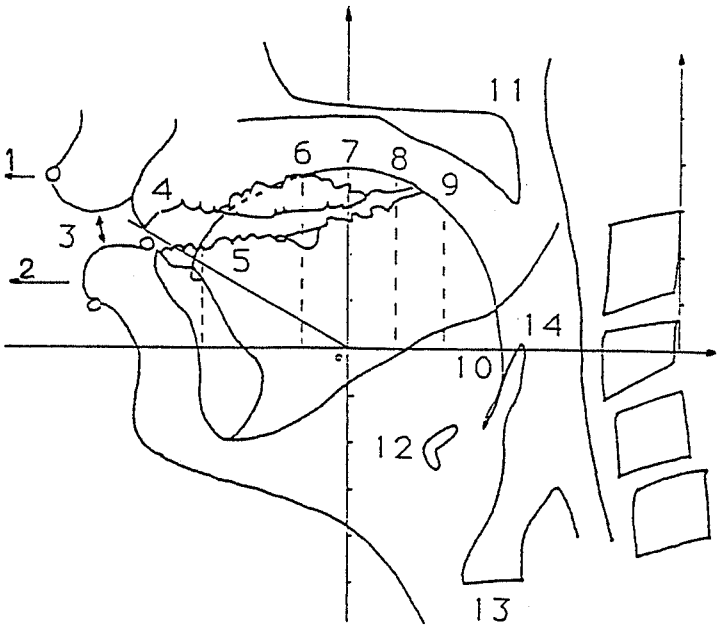


Fig. 1 : Profil radiologique du conduit vocal, paramètres mesurés.

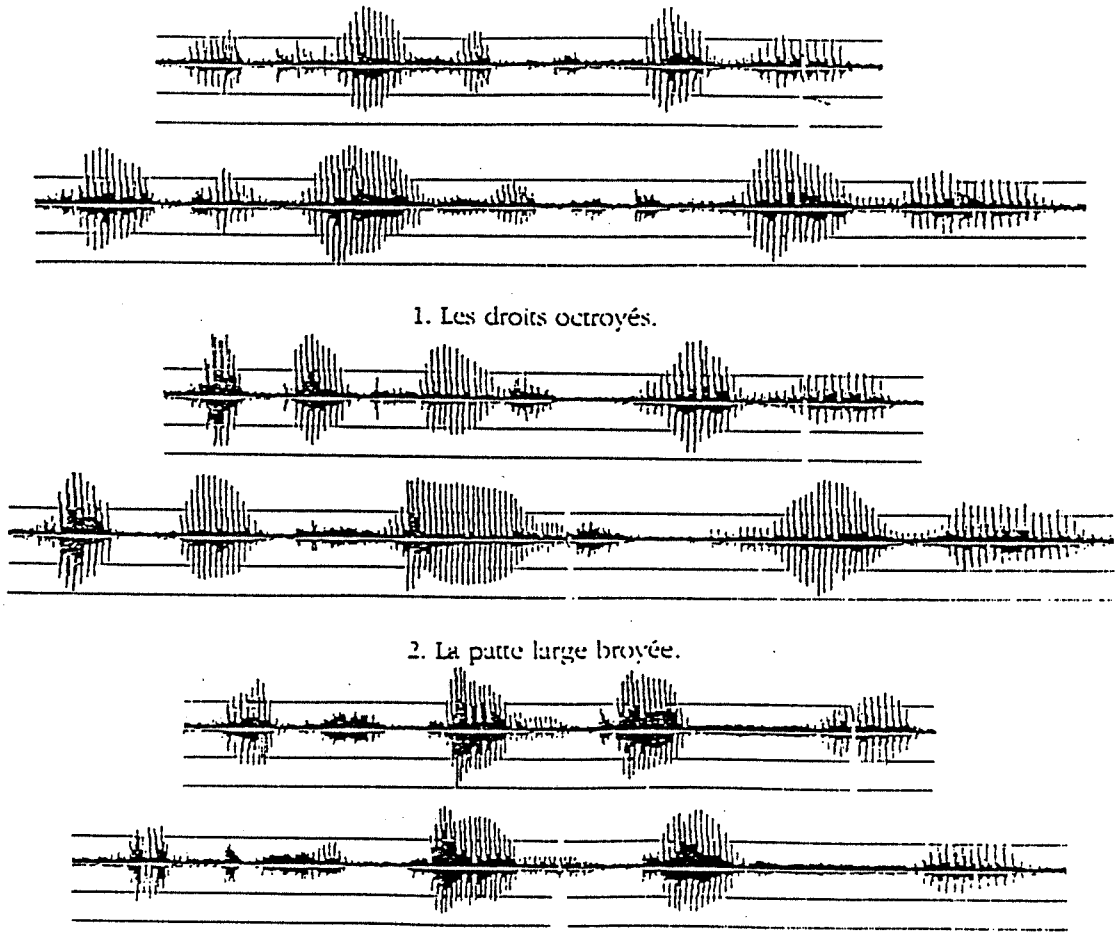


Fig. 2 : Oscillogrammes, comparaison DL/DR.

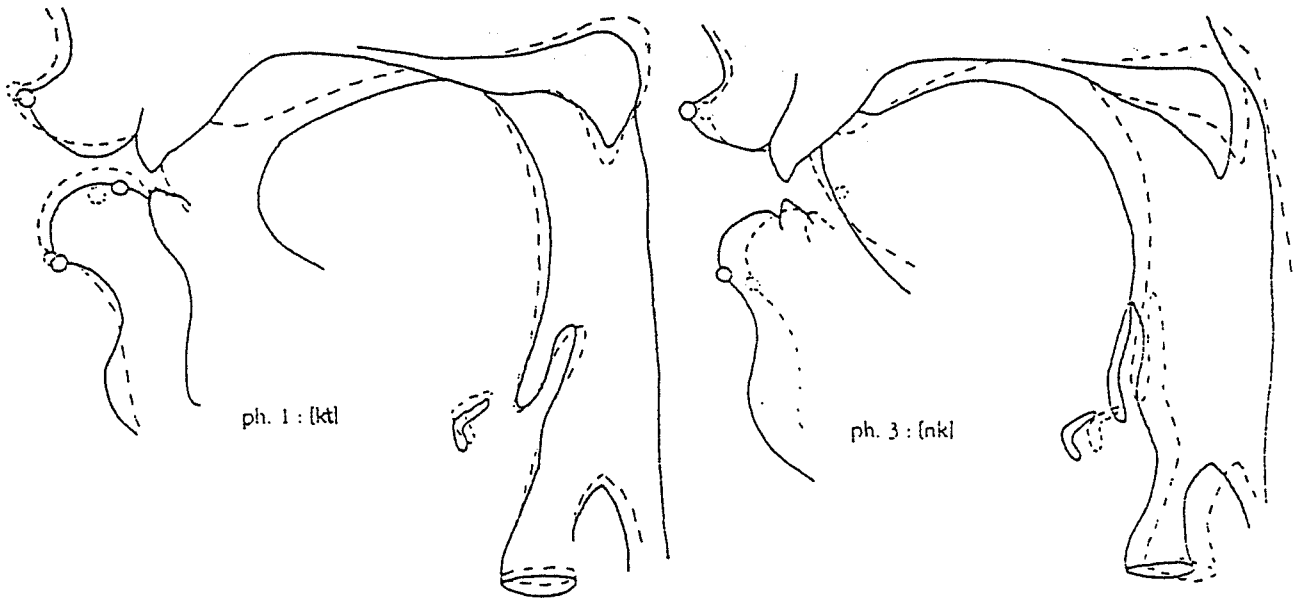


Fig. 3 : Vues de profil.
Trait plein : DL
Trait pointillé : DR

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION AFFICHEE : A4

PHONETIQUE ET PROSODIE

THE PHONETIC TYPOLOGY OF GERMANIC LONG AND SHORT VOWELS

R.K.POTAPOVA
STATE LINGUISTIC UNIVERSITY,
MOSCOW, RUSSIA

Résumé

A review of current typological research on the domain of the perception and instrumental analysis of long (tense) and short (lax) vowels in the Germanic languages (German, English, Swedish, Norwegian, Dutch) is given. It is shown that the phonological opposition "long-short" (tense-lax) for Germanic vowels has not quantitative, but qualitative correlates.

Key words: vowel, long, short, typology

INTRODUCTION

The term "typology" has a number of different uses both within linguistics and without. The common definition of the term is roughly synonym with "taxonomy" or "classification", the classification of phenomena under study of types particularly structural types. The broadest and most unassuming linguistic definition of "typology" refers to a classification of structural types across languages. In this case typology has to do with cross-linguistic comparison of some sort.

A more specific definition of "typology" is that it is, the study of linguistic patterns that are found cross-linguistically, in particular, patterns that can be discovered solely by cross-linguistic comparison. The classic example of typology under this definition is the implicational universal.

There is a still specific of "typology". On this view, typology represents an approach to the study of language that contrasts with prior approaches. In this definition "typology" is an approach to linguistic theorizing, or more precisely a methodology of linguistic analysis, that gives to different kinds of linguistic theories than other approaches.

We will use the term "typology" to refer to the subdiscipline: cross-linguistic comparison of segmentation phenomena. The primary purpose of this paper to discuss the kinds of cross-linguistic patterns that have been discovered and methodological and empirical issues raised by the study of these patterns.

Typological classification is the process of describing the various linguistic types found across languages for some parameter, such as a nature and number of speech segments. The usual procedure for initiating a cross-linguistic comparison of phonetic phenomena for the purposes of a typological analysis is to survey the range of structures used for the phenomena in question. These structures (in this case perceptional speech segments) are called types or strategies. Also, the typology of art of auditory segmentation of speech displays practically the full range of segmentation strategies.

The process of segmentation of speech image is a basis of our typological description. It is necessary to distinguish three kinds of speech segments: production, perception and acoustic segments. In all this cases the process of speech segmentation includes simultaneously two phenomena: continuity and discontinuity. The aim of perceptional segmentation could be characterized as description of speech continuum with special symbols with correspondence to a number and perceptional characteristics of timbre segments.

ANALYSIS

The present research is the first comparative experimental phonetic study of the segmental structuring of vowels in Germanic languages, involving statistically reliable material and a complex method, providing for a perceptional and acoustic with reference to spectral analysis, as well as mathematical and statistical types of analysis.

The experimental material as applied to German, English, Swedish, Danish, Norwegian and Dutch compiled according to the principle of interlanguage minimal pairs meeting as far as possible the requirements of the identical sound content, accentual and rhythmic structure and the segmental length of the utterance.

The hypothesis of the experiment was formulated as follows: the auditory identification of long and short (tense and lax) vowels in the Germanic languages and consequently the type of a release of vowel are based not on the vowel duration, but on the vowel qualitative dynamics. In other words, the character of the connection of the vowel with the following consonant is determined by the qualitative dynamics of the preceding vowel and by the specific character of its timbre structural organization. The method of auditive segmentation was used in the experiment.

The findings obtained suggest that the perception of absolute duration of vowels in the Germanic languages does not depend only on quantitative characteristic of these speech segments.

The presented stimulus is identified as short when the vowel nucleus is positioned in the center of the sound, the number of timbre segments is mini-

mal (n=2-3). The change of timbre segments modifies from more open speech segment to more closed ones.

Quite different picture is observed with the increase in the number segments (n=4-6), the sound nucleus is positioned in the release of the sound and timbre segments changing from more closed to more open ones are identified as long.

Subjectively (perceptional) the nucleus timbre segment is singled out as a segment corresponding with the typical variant of the given phoneme, but objectively (instrumental) this segment has the characteristics which are close to the F_n - structure of the given segment.

The results of the performed auditive and spectral analysis make it possible to differentiate Germanic vowels with the early qualitative markedness and localization of the nucleus timbre segment in the final part of the sound which is characteristic of the so-called long vowels, vowels with a later stage qualitative markedness and the localization of the timbre nucleus segment in the center of the so-called short vowels (See Potapova: 1986).

For example, auditive and acoustic data for "long" (tense) vowels:

[a:] in Swedish includes a set of timbre segments with following F_n - structure:
a_1 (F_1 - 650 Hz; F_2 - 1000 Hz), a_2 (F_1 - 650 Hz; F_2 - 900 Hz), a_3 (F_1 - 750 Hz; F_2 - 1100 Hz), a_4 (F_1 - 550 Hz; F_2 - 1000 Hz).

[a:] in German includes a set of timbre segments:
a_1 (F_1 - 750 Hz; F_2 - 1300 Hz), a_2 (F_1 - 650 Hz; F_2 - 1100 Hz), a_3 (F_1 - 750 Hz; F_2 - 1000 Hz), a_4 (F_1 - 400 Hz; F_2 - 1000 Hz).

[a:] in Dutch has a set of timbre segments too:
a_1 (F_1 - 750 Hz; F_2 - 1300 Hz), a_2 (F_1 - 650 Hz; F_2 - 1200 Hz), a_3 (F_1 - 750 Hz; F_2 - 1100 Hz), a_4 (F_1 - 750 Hz; F_2 - 1100 Hz).

For example, auditive and acoustic data for "short" (lax) vowels:

[a] in Swedish is characterized by following timbre segments:
a_1 (F_1 - 500 Hz; F_2 - 1600 Hz), a_2 (F_1 - 500 Hz; F_2 - 1400 Hz), a_3 (F_1 - 400 Hz; F_2 - 900 Hz).

[a] in German includes following timbre segments:
a_1 (F_1 - 600 Hz; F_2 - 1800 Hz), a_2 (F_1 - 750 Hz; F_2 - 1200 Hz).

[a] in Dutch is characterized by following timbre segments:
a_1 (F_1 - 750 Hz; F_2 - 1000 Hz), a_2 (F_1 - 750 Hz; F_2 - 1100 Hz).

The analysis of the intensity curve, its development in time, the angle of its decrement indicates that short and long vowels being different in the number of intensity peaks practically have no difference in the off-glide vowel phase in the transition to a post-vocalic consonant. Our experiments show that phonological quantitative opposition of vowels in Germanic languages has not quantitative but qualitative speech basis first of all in the flow of speech.

CONCLUSIONS

Linear character of speech continuum presupposes realization of speech segments in dimensions and organized in time. Such realization of speech segments is impossible without the coexistence of phenomena of continuity and discontinuity. Specific nature of passing from a continuous character of speech production to a discrete one depends on universal physiological factors, the basis of pronunciation, phonological peculiarities of language, concrete conditions of speech act. In this connection speech, as any other system, is a complex whole, and its constituent parts interact according to the principle of multi-level stratification (See Potapova: 1991).

From this point of view it would be necessary to define and to describe all types of speech segments and its classification for different languages, the groups of languages with identical features of segments, similar features and unsimilar ones of different languages.

The description of speech segment system of different languages on the domain of speech production, perception and acoustics, the classification of ingredients of this systems and subsystems, the kinds of corresponding relations between them are very important for the general theory of language and for aims of applied linguistics: speech analysis and synthesis, automatic speech recognition, language teaching, learning and acquisition.

In this case there is opportunity to find and to describe types of segmentation, its articulatory, perceptional and acoustic characteristics, to classify this data and to define some universals on this domain.

REFERENCES

Potapova, R. (1986), Slogovaja phonetika germanckich jazykov, Moscow: Vysshaja shkola, 144 p.

Potapova, R. (1991), The typology of speech segment units, Proc. of the XIIth Intern.Congress of Phonetic Sciences, Aix-en-Provence, 1991, 166-169.

Potapova, R. (1992), The auditory identification of long and short vowels in the Germanic languages, Abst. of the XVth International Congress of Linguists, Laval, Canada, 1992, 8.

LA SEGMENTATION DE LA PAROLE CHEZ DES ADULTES LETTRÉS ET ILLETTRÉS ARABOPHONES: APPROCHE PSYCHOLINGUISTIQUE

FARID, Mohamed
U. F. R. L. UNIVERSITE PARIS 7
LABORATOIRE DE PHONETIQUE

Résumé

Une expérience est menée pour étudier l'effet de la lecture sur le développement de la capacité à traiter la parole en segments. Les résultats confirment l'hypothèse selon laquelle la conscience phonologique ou segmentale est plus élevée chez le groupe des lettrés que chez le groupe des illettrés qui sont tous arabophones du Maroc. Cette contribution est inspirée du travail de Morais, Bertelson, Cary et Content (1986).

Mots-clés: Segmentation, illettrisme, arabe, conscience phonologique.

INTRODUCTION

La segmentation de la parole en tant que processus cognitif a été beaucoup étudiée par les phonéticiens, les linguistes et les psycholinguistes. Les quinze dernières années ont vu le développement d'un réseau de recherches portant sur la capacité à segmenter un signal de parole en ses unités composantes. Les psycholinguistes appellent cette capacité "la conscience phonologique". Selon Morais (1991), la conscience phonologique réfère à une catégorie spéciale des représentations phonologiques. Elle est une sorte de connaissance phonologique mais elle diffère de celle-ci, qui est utilisée dans la production et la compréhension du langage, par le fait qu'elle réfère à des représentations conscientes des propriétés et des constituants phonologiques de la parole. En outre, il faut souligner que la conscience phonologique est logiquement liée à l'acquisition de la lecture et de l'écriture dans un système alphabétique. Morais (1985) souligne que la conscience du phone est une composante cruciale de la lecture et de l'écriture. Son développement n'est permis qu'avec l'apprentissage de ces deux dernières compétences.

Il est à noter qu'il existe quatre niveaux de conscience phonologique: conscience des mots, conscience des syllabes et conscience des unités infra-syllabiques (attaque et rime).

Pour appuyer la thèse de l'existence de la conscience phonologique, certaines études psycholinguistiques ont montré que les enfants à l'étape préscolaire, comme les illettrés, sont incapables de manipuler des segments ou des unités infra-morphémiques. Leur performance est faible en ce qui concerne les phonèmes mais assez bonne en ce qui concerne les syllabes.
Je cite deux types d'études concernant l'examen de l'hypothèse de la conscience phonologique:
-Expériences d'entraînement à la conscience phonologique (l'emploi de tests dits métaphonologiques).
- Etudes longitudinales ou de corrélation.
Morais, Cary, Alegria et Bertelson (1979) ont montré que les illettrés adultes (portugais) ne pouvaient pas analyser explicitement la parole en tant que suite de phones. Ceci confirme leur hypothèse que la conscience de la structure segmentale de la parole n'émerge pas spontanément, par simple maturation cognitive, mais apparaît généralement lorsqu'on est confronté à la situation d'apprentissage de la lecture. Soulignons que les tâches utilisées dans l'étude de Morais et al. (1979), consistaient soit à soustraire, soit à additionner, un phone consonantique au début d'un mot. Les résultats obtenus chez les deux groupes sont 19 % de réponses correctes en moyenne pour les illettrés et 73 % pour les ex-illettrés. Ces résultats montrent clairement que les adultes illettrés ne se représentent pas la parole comme une séquence de phones.

Dans une autre étude, Morais et al. (1986) ont posé que l'alphabétisation a des effets spécifiques sur la capacité d'analyse segmentale de la parole et la reconnaissance des mots. Les travaux de Morais et al; (1979, 1986) effectués au Portugal sur des sujets adultes illettrés et des sujets alphabétisés à l'âge adulte ont conduit à l'idée que la capacité à analyser intentionnellement la parole en phonèmes ne s'acquiert généralement pas indépendamment de l'acquisition du du code alphabétique (voir également Read, Zhang, Nie et ding, 1986).
Morais et al. (1986) ont soumis à leurs illettrés et ex-illettrés portugais une série de tâches métaphonologiques (comme la segmentaion de la parole par suppression ou addition de segments phoniques, détection de phones et de rimes.

Les résultats obtenus révèlent que les ex-illettrés, c'est-à-dire des personnes de même âge et de même origine sociale que les illettrés, dépourvus aussi d'un passé scolaire normal mais ayant appris à lire et à écrire à l'âge adulte, obtiennent une performance relativement bonne dans les tâches d'analyse phonémique intentionnelle et peuvent donc vraisemblablement être crédités d'un niveau non négligeable de conscience des phonèmes (Morais et al., 1979, 1986). Les auteurs concluent leurs études par l'idée selon laquelle les illettrés ne possèdent pas généralement la capacité d'analyse intentionnelle de la parole en phonèmes.
Morais et al. (1986) constatent qu'il y a une corrélation forte entre l'acquisition d'une lecture alphabétique et la capacité à traiter la parole au niveau des unités infra-morphémiques comme la syllabe et le phonème.

La présente étude se veut une contribution à l'étude de quelques aspects psycholinguistiques de la segmentation de la parole en unités minimales chez des adultes lettrés et illetrés arabophones vivant en France.
Je propose l'hypothèse selon laquelle l'apprentissage ou l'acquisition de la lecture dans un système d'écriture donné, essentiellement alphabétique, a un effet sur la performance des sujets à analyser et segmenter la parole en unités syllabiques et infra-syllabiques (Farid, 1991). Le bon lecteur acquiert, par ce fait, une conscience phonologique lui permettant d'analyser, selon un traitement conscient, la structure phonémique des mots qu'on lui présente.
Généralement, les illettrés vivant dans un milieu socioculturellement pauvre ne peuvent pas développer des capacités intentionnelles leur permettant de répondre facilement à certaines tâches d'analyse segmentale. Je précise ce point car j'ai constaté que certains de mes illettrés, vivant en France, ont une performance assez bonne en segmentation progressive. Je pourrais expliquer cela par l'influence du français sur les processus psycholinguistiques de traitement de l'arabe; c'est ce que j'appelle *"effet du bilinguisme"*. En ce sens, le bilinguisme pourrait contribuer au développement de la conscience métalinguistique chez le sujet bilingue. Autrement dit, on pourrait parler d'un *"export"* des stratégies de segmentation du français pour les appliquer au traitement segmental de l'arabe.

EXPERIENCE

METHODE

Sujets: Deux catégories de sujets ont participé à l'expérience de segmentation progressive de la parole. La catégorie des illettrés est représentée par un groupe composé de huit sujets, deux femmes et six hommes. Ce sont tous des immigrés Marocains vivant à Paris et dans sa région depuis plusieurs années.Ils sont âgés de 30 à 65 ans. Ils communiquent assez difficilement en français. Ils sont d'origine paysanne. La deuxième catégorie comprend deux groupes de lecteurs dont 10 bons lecteurs et 8 lecteurs moyens .Les bons lecteurs sont tous des étudiants universitaires Marocains âgés de 22 à 31 ans dont deux filles et huit garçons. Tous ces sujets ont reçu au Maroc une formation bilingue solide en arabe standard et en français.
Les lecteurs moyens sont 3 femmes et 5 hommes âgés de 21 à 51 ans. Les trois femmes ont un niveau du CMII marocain, et lisent en arabe et en français. Les hommes sont des immigrés ouvriers; l'un d'eux est ex-illettré, il n'a jamais été à l'école à l'âge scolaire. Il a reçu des cours d'alphabétisation en arabe dans des associations islamiques et en français dans un centre municipal. Il sait lire et écrire en arabe mieux qu'en français. Les autres sujets ont arrêté leur école au niveau du CMII marocain. Ils gardent encore des notions de lecture et d'écriture.

Matériel: A l'inverse de Morais (1986), j'ai construit 10 phrases-cibles d'une moyenne de 7, 6 mots par phrase. Les mots de ces phrases appartiennent à l'arabe marocain parlé. Le plus long mot dans ces phrases est constitué de 3 syllabes. Les disyllabes sont les plus fréquents du genre [CV-CV] (ex: **[daru]** "ils ont fait"). Les plus petits mots sont des morphèmes grammaticaux monosyllabiques de structure CV (ex: **[wa]** "et").

Tâche et procédure: Les stimuli (10 phrases) ont été enregistrés, dans un premier temps, sur une bande magnétique puis copiés sur une cassette. Avant de commencer le test, j'ai proposé aux sujets quelques essais afin qu'ils puissent assimiler la tâche.
Directives:
- Ecoutez attentivement la phrase qui sera prononcée.
- Essayez de la mémoriser entièrement si c'est possible.
- Répétez ce que vous venez d'entendre.
- Dites un morceau de cette phrase.
- Ensuite, dites une partie de ce morceau jusqu'à ce que vous ne puissiez plus réduire davantage.

Les sujets avaient pour tâche de segmenter progressivement les phrases qui leur étaient exposées auditivement.

RESULTATS

Unités Isolées	Illettrés	L.M.	B.L.
G.M.	20, 91	35, 17	43, 92
2Mots	27, 40	31, 51	41, 10
1 Mot	32, 54	33, 90	33, 56
Syllabe	25, 08	26, 76	48, 16
Phone	23, 12	26, 63	50, 25

* L. M. : Lecteurs moyens
B. L. : Bons Lecteurs
G. M. : Groupe de Mots

Tableau. Pourcentages moyens de segmentation progressive par isolement d'unités linguistiques chez les illettrés, les bons lecteurs et les lecteurs moyens arabophones.

Les taux moyens de segmentation ont été obtenus à partir du nombre de réponses (ou segmentations) pour chaque énoncé et chez chaque sujet.

Les résultats du tableau montrent que les illettrés ont une performance variée concernant l'isolement des unités linguistiques lors de la segmentation progressive. C'est au niveau du mot qu'ils ont la meilleure performance. Les pourcentages moyens d'isolement du phone (23, 12) et de la syllabe (25, 08) sont différents: t = 1, 848; dl= 9; p< 0, 005. C'est aussi le cas de la différence entre l'isolement du phone et celui du mot: t= 6, 206; dl= 9;
p< 0, 001.
La performance des lecteurs moyens n'est pas très différente dans les 5 types d'isolement. Nous avons un taux d'isolement allant de 26,63 % pour un phone, à 35,17 % pour l'isolement de plusieurs mots. Ce dernier résultat montre clairement que les lecteurs moyens ont une bonne performance en ce qui concerne la mémorisation du matériel verbal ou linguistique. Je note également que les sujets ont la même capacité à isoler les phones que les syllabes: t= 0,777; dl= 9; N.S. Cependant, la différence moyenne entre l'isolement du phone et celui du mot est significative: t= 2,165; dl= 9: p< 0, 03. En effet, c'est à partir de l'unité mot que la performance des sujets devient importante.
Je constate également d'après ce tableau que les bons lecteurs sont très performants sur les 5 types d'isolement d'unités segmentales. Plus les bons lecteurs s'éloignent du mot, plus leur performance s'élève. C'est au niveau du mot que ces sujets ont la même performance que les deux groupes précédents (32, 54 %, 33, 9 %, 33, 56 %). Cela va dans le sens de l'hypothèse de départ selon laquelle les bons lecteurs ont la capacité d'atteindre les unités syllabiques et phonémiques dans leur segmentation progressive des phrases.
Le mot représente sans doute l'unité lexicale la plus aisément traitée par l'ensemble des sujets. Une analyse de variance (ANOVA) a montré qu'il y a un effet significatif de l'alphabétisation: $F(4,10)=10,84$; $p<0,005$

DISCUSSION

Dans une langue comme l'arabe, il est difficile, voire impossible de segmenter une syllabe en consonnes et voyelles car dans cet alphabet, les voyelles (u, a, i) n'ont pas un statut de graphèmes comme les consonnes. Elles sont représentées graphiquement par des marques diacritiques. Donc, la syllabe [ka] en arabe correspond graphiquement à un temps unique, à l'inverse du français dans lequel la syllabe correspond à l'addition de deux temps: **t1 + t2 ; (k= t1), (a = t2).**
En d'autres termes, la syllabe écrite, en arabe, est constituée d'une consonne et d'une marque diacritique représentant la voyelle qu'il est impossible d'isoler toute seule.
En conclusion, les bons lecteurs manifestent une conscience phonologique plus grande que celle des autres sujets car ils ont la capacité d'isoler les unités infra-lexicales constituant la structure phonémique des mots et des phrases. Le développement de cette conscience est assuré par leur pratique de la lecture et de l'écriture .Ces deux systèmes cognitifs favorisent chez le lecteur-scripteur la capacité à manipuler les éléments ou unités de la parole. La manipulation de la structure segmentale des mots peut résulter d'un traitement conscient ou intentionnel des éléments de la parole. Ainsi, segmenter d'une manière progressive une phrase exige du sujet une certaine stratégie de segmentation lui permettant de mémoriser d'abord le matériel verbal, ensuite de le traiter en fonction de ses connaissances linguistique, métalinguistique et extra-linguistique. L'interaction des illettrés avec un milieu socioculturel riche leur permet de développer certaines capacités de traitement segmental de la parole. Cette proposition a été confirmée par les résultats de certains illettrés dans la tâche de la segmentation progressive des phrases. Je veux dire par là que la situation bilingue peut avoir un effet positif sur le traitement perceptif et segmental de la parole. Autrement dit, on peut parler d'"effet d'interférence" sur la segmentation de la parole.
L'essentiel de l'analyse de la segmentation progressive des énoncés va dans le sens de l'hypothèse avancée par l'équipe de Morais.

CONCLUSION GENERALE

Le présent travail a été consacré à l'étude de la performance des sujets lettrés et illetrés lors de la segmentation progressive des énoncés de l'arabe marocain. L'objectif était de vérifier l'hypothèse selon laquelle l'acquisition de la lecture et de l'écriture développe chez le sujet la capacité à traiter la stucture segmentale des mots parlés. Cette hypothèse est inspirée du travail réalisé par Morais et al. (1986).

D'après les résultats obtenus, il a été confirmé que la lecture a bien un effet sur la capacité à segmenter la parole. Mais les résultats de certains illettrés de la présente expérience m'ont amené à proposé une hypothèse qui prédit que leur statut de bilingue pourrait jouer un rôle dans le développement de la capacité à manipuler les segments de la parole, surtout chez des sujets illettrés.
Enfin, on peut dire que l'étude des effets de l'alphabétisation sur la reconnaissance et la segmentation des mots parlés peut par conséquent contribuer à une meilleure compréhension de l'architecture du système de traitement du langage.

Remerciements. Je tiens à remercier vivement Mademoiselle Isabelle Malderez et Messieurs Georges Boulakia et Jean-Yves Dommergues d'avoir revu et corrigé une version de ce travail.

BIBLIOGRAPHIE

Farid M.. (1991). *Quelques Aspects de la segmentation et de la perception de la parole chez des sujets lettrés et illettrés adultes: Etude psycholinguistique.* Mémoire de D. E. A. en Phonétique, Université Paris VII.

Morais J. (1985). Literacy and awareness of the units of speech: implications for research on the units of perception. *Linguistics,* **23,** 707-721.

Morais J. (1991). Phonological Awareness: A Bridge Between Language and Literacy.. In D.J. Sawyer & B.J. Fox (Eds.) *Phonological Awareness in Reading: The Evolution of Current Perspectives.* Springer-Verlag.

Morais J., Bertelson, P., Cary, L., & Content, A. (1986). Literacy training and speech segmentation. *Cognition,* **24,** 45-64

Morais J., Cary L., Alegria J., & Bertelson P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition,* **7,** 323-331.

Read, C., Zhang Y., Nie H., Ding B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic spelling. *Cognition,* **25,** 21-52.

Annexe: Matériel utilisé dans l'expérience

1- nnas qeddam babu kaytkelmu
(les gens discutent devant sa maison)

2- xeSna daba n$a&lu DDaw Hit TaH Dlam
(la nuit tombe, on doit allumer la lumière)

3- &emmer kasu Hlib ba$ yefTer
(il a rempli un verre de lait pour déjeuner)

4- snanu kulhum TaHu mabqa yeqder yakul mezyan
(il a perdu toutes ses dents, il ne peut plus manger bien)

5- had rrajel krim kay Seddeq ssukkar welxubz &la lmasakin
(cet homme est généreux, il donne la charité aux mendiants)

6- had lla&ib faz belkura ddahabiyya &la le&bu lmumtaz
(cet joueur a obtetenu un ballon d'or pour son excellent jeu)

7- &endu susa ffummu matxallih$ yakul &la xaTru
(il a une carie qu'elle ne le laissse pas manger à l'aise)

8- mat$urbu$ men had lma rah bared bezzaf
(ne buvez pas cette eau, elle très froide)

9- jabu SSiyyada fila wqerd men ghaba b&ida
(les chasseurs ont ramené , d'une forêt lointaine, un éléphant et un singe)

10- waxxa had lmra zina hetta waHed mabghaha
(même si cette femme est belle personne ne l'adore)

CARACTERES SPECIFIQUES DES VOYELLES NASALES DU FRANCAIS

Jean-Sylvain LIENARD*
Maria-Gabriella DI BENEDETTO**

* LIMSI-CNRS BP133 91403 ORSAY Cedex
** INFOCOM Università La Sapienza, Via Eudossiana 18, 00184 ROMA

Résumé

Dans cette étude on cherche à caractériser les voyelles nasales du français dans la perspective d'aider à l'apprentissage de leur prononciation par un non-francophone. Dans un premier temps on dresse un panorama des connaissances disponibles sur la nasalité vocalique, selon divers points de vue : production, acoustique, perception et synthèse. Dans un deuxième temps on décrit la base de données utilisée, comportant des voyelles isolées émises par divers locuteurs. Puis on décrit les traitements permettant d'associer à chaque segment une enveloppe spectrale. Chaque nasale est ensuite comparée visuellement avec l'orale correspondante, en vue de faire ressortir leurs ressemblances et différences. Ces observations sont confrontées aux connaissances publiées. Enfin on décrit une expérience de reconnaissance automatique, dans laquelle la nasale produite par chaque locuteur dans diverses conditions d'élocution est comparée à sa reconstruction à partir des orales de la même série.

INTRODUCTION

Cette étude préliminaire a été menée dans le but de permettre au système SPELL d'aider un élève à apprendre la prononciation des voyelles nasales du français[1]. On suppose qu'il a acquis auparavant la prononciation des voyelles orales. Le système doit décider si la nasale produite par l'élève est correcte, sans pouvoir se référer à un modèle provenant du même locuteur. Cette question suppose que les connaissances concernant la nasalité vocalique pour le français soient disponibles et utilisables en reconnaissance automatique. Nous commencerons par faire le point sur la nasalité vocalique, selon divers points de vue : production, analyse, perception/synthèse, reconnaissance. Puis nous décrirons nos données et l'analyse qui en a été faite afin de mettre en évidence l'enveloppe spectrale. Nous montrerons que, s'il est difficile de dégager des indices de la nasalité dans l'absolu, par contre on peut caractériser acoustiquement chaque nasale par rapport à certaines voyelles orales du même locuteur dans les mêmes conditions. Enfin nous montrerons comment ces indices peuvent être validés dans une expérience de reconnaissance automatique.

[1] Ce travail a été effectué en partie dans le cadre du contrat ESPRIT SPELL-II (projet 7153), dont le partenaire principal est la société AGORA CONSEIL, en partie dans le cadre d'une coopération entre le CNRS et le CNR (Consiglio Nazionale delle Ricerche).

LA NASALITE VOCALIQUE

Production

La nasalité est produite par la mise en parallèle du conduit nasal avec le conduit oral, par abaissement du velum. Le couplage avec le conduit nasal entraîne l'apparition de plusieurs résonances et antirésonances (pour une discussion nous renvoyons le lecteur à Maeda [1] et Castelli [2]). Mentionnons ici les principaux effets de la nasalisation établis par calcul :

- Premier formant nasal, Fn1, aux alentours de 250 Hz. Cette résonance notée par de nombreux auteurs, semble se comporter comme une résonance de Helmholtz et non comme une fréquence propre d'un tuyau acoustique.
- Série de formants aux environs de 1000, 2000, 3000 et 4000 Hz. Ces formants correspondent aux fréquences propres du conduit nasopharyngal considéré comme un tuyau 1/2 onde .
- Paire pôle-zéro dans la région de F1 : en particulier les simulations montrent qu'une paire pôle-zéro apparaît entre 300 et 1000 Hz dès l'ouverture du velum, même pour de petites aires de couplage.
- Amortissement général du spectre : les pôles du conduit nasopharyngal, les paires pôle-zéro dues au couplage et la dissipation d'énergie dans les fosses nasales ont pour effet d'amortir le spectre dans son ensemble.

Analyse

Si l'on se place du point de vue de l'analyse, on part de matériaux réels (voyelles émises par des sujets bien caractérisés, au lieu de calculs sur un conduit vocal idéalisé), dans lesquels on cherche des indices significatifs du point de vue de la phonétique, de la production ou de la perception.

Le travail de F.Lonchamp en 1978 [4] a été mené dans cette perspective. Deux locuteurs masculins (phonéticiens spécialistes) ont enregistré chacun 6

exemplaires des quatre voyelles nasales du français, dans des conditions de laboratoire (voyelles parfaitement stables, Fo fixé à 120 Hz). Chaque exemplaire a fait l'objet d'une analyse spectrographique, et l'auteur a déterminé manuellement la position des divers maxima spectraux à partir des "sections" sonagraphiques (spectre en bande étroite, 45 Hz). Il a ensuite confronté la position de ces divers maxima d'une part avec les mesures publiées par d'autres auteurs, et d'autre part avec les prédictions des divers spécialistes de la production. Il a complété son étude par des synthèses au moyen d'un synthétiseur à formants de type série. Voici quelques-unes de ses conclusions :

- présence d'un premier formant nasal Fn1 vers 280 Hz,
- présence d'un second formant nasal Fn2 vers 2000 Hz,
- les voyelles nasales semblent caractérisées par la présence de deux "formants oraux décalés" F'1 et F'2, sans qu'il soit besoin de faire appel à un indice général de nasalité,
- la voyelle orale française la plus voisine de [ɑ̃] est plutôt [ɔ] que [a],

Perception

La plupart des études sur la perception de la nasalité vocalique ont été menées au moyen de la synthèse. Dès 1954 P.Delattre [5] a signalé plusieurs indices utiles pour synthétiser une voyelle nasale à partir de la voyelle orale correspondante, soit par ordre d'importance :

- l'affaiblissement du premier formant F1,
- l'accentuation de la résonance basse à 250 Hz (premier formant nasal),
- la mise en place d'un second formant nasal vers 2000 Hz.

Mais le travail le plus significatif dans la méthodologie synthèse-perception est sans doute celui de S.Hawkins et K.Stevens en 1985 [6]. Utilisant un synthétiseur de Klatt en configuration série ils ont synthétisé divers intermédiaires entre une voyelle orale et la nasale correspondante, en faisant varier les caractéristiques d'un pôle et d'un zéro placés dans la région de F1. Ce traitement a été appliqué à cinq voyelles différentes ([i], [e], [a], [o], [u]). Les synthèses ont été évaluées par divers groupes d'auditeurs, notamment par un groupe d'Américains pour qui l'opposition oral / nasal n'est pas phonologiquement pertinente, et divers groupes de locuteurs dont la langue maternelle possède cette opposition pour toutes ou plusieurs des cinq voyelles étudiées. Pour chacune des voyelles traitées les auteurs ont pu trouver un positionnement optimum du pôle et du zéro additionnels, tel que la nasalité a été correctement perçue par tous les groupes d'auditeurs.

Reconnaissance

Il existe peu d'études sur la reconnaissance automatique de la nasalité vocalique. Mentionnons les travaux de Maeda [1], portant sur des voyelles nasales synthétiques produites par son modèle phonatoire, et de Glass [3], portant sur la nasalisation vocalique en anglais américain. Ce dernier a obtenu un taux d'erreur de l'ordre de 26% sur un corpus multilocuteur, la décision à prendre étant binaire (la voyelle étudiée était soit purement orale, soit nasalisée au contact d'une consonne nasale : le choix au hasard donnerait un taux d'erreur de 50%).

CORPUS ET ANALYSES

Le corpus Corenc a été conçu avec l'idée d'inclure plusieurs sources de variabilité dans les données et d'étudier leurs interactions. Il s'agit d'une petite base de données, comprenant 60 séries de 12 voyelles françaises, prononcées par 13 locuteurs en plusieurs sessions, avec diverses forces de voix. Dans cette phase préliminaire on a choisi de traiter des voyelles isolées et d'utiliser des locuteurs ordinaires, non-spécialistes de phonétique.

Dans l'inventaire des voyelles utilisées on a laissé de côté le [ɔ] qui, à l'inverse des autres voyelles du français, ne peut être noté sous forme d'un mot lexical, ainsi que le [œ̃], qui n'est plus phonologiquement distinct de [ɛ̃] et le [ə] qui n'a pas d'existence autonome. On a aussi considéré un seul [a], le [a] moyen qui correspond aujourd'hui à la majorité des réalisations en français standard. On a demandé au locuteur de répéter les voyelles prononcées par l'opérateur, avec une force de voix en rapport avec la distance à laquelle se trouve ce dernier (30 cm, 1,50 m et 7m). On a évité de se placer dans les cas extrêmes (voix chuchotée et voix criée), pour rester dans le registre des variations habituelles et presque inconscientes de la voix. Les locuteurs sont 6 hommes et 7 femmes, adultes, originaires pour la moitié d'entre eux de la région de Grenoble. Les enregistrements ont été faits en deux sessions distantes de six mois ; certains locuteurs ont participé aux deux sessions.

Les 720 voyelles ont été évaluées auditivement par un groupe de 5 auditeurs majoritairement parisiens et distinct du groupe des locuteurs [7]. Les conclusions en ce qui concerne le problème traité ici sont les suivantes :

- les voyelles isolées sont très largement identifiées par les auditeurs (environ 9% d'erreur, ce qui est relativement faible),
- les voyelles nasales sont moins bien reconnues que les voyelles orales (environ 15% d'erreur, contre 7%, la nasale la plus difficile à identifier étant "on", avec 23% d'erreur),
- les confusions sont fortement liées au locuteur : certains locuteurs semblent avoir plus de mal que d'autres à articuler correctement les nasales ou certaines nasales.

L'analyse a été faite dans la perspective d'une reconstitution ultérieure du signal par une méthode dérivée de l'analyse sinusoïdale [8][9][10]. Dans un premier temps on recherche pour chaque segment l'instant qui concilie le plus fort niveau sonore avec la plus forte stabilité spectrale. Dans la suite chaque voyelle sera représentée par le spectre prélevé à cet instant du segment, avec les paramètres suivants : durée 51,2 ms, fenêtre de Hamming, amplitudes en dB. Ce spectre est ensuite réduit à ses pics, après une correction d'amplitude (interpolation parabolique).

Le spectre brut fait apparaître de nombreux pics parasites, qui très souvent ne correspondent pas à des détails spectraux audibles. Pour les éliminer on calcule en chaque point du spectre brut une "fonction de masquage", parabole convexe centrée sur chaque pic et de largeur (évaluée à -20 dB) variant entre 120 Hz dans le bas du spectre et 330 Hz dans le haut. En chaque point du spectre on ne retient que le maximum de toutes les fonctions de masquage concernant ce point, ce qui constitue le "spectre masqué", auquel on fait ensuite subir un léger lissage. Le résultat de cette opération est un élargissement des raies spectrales et un déplacement d'ensemble du spectre vers les valeurs maxima ; dans ce processus les pointes de faible intensité par rapport à leurs voisines sont masquées. Pour obtenir l'enveloppe il suffit alors de ne retenir que les pics du spectre brut dépassant la valeur du spectre masqué lissé, et d'interpoler linéairement entre pics successifs.

L'échelle linéaire de fréquence donne une trop grande importance à la partie haute du spectre. L'utilisation d'une échelle Bark selon une formulation dûe à Schroeder [11] apporte une correction à cette distorsion.

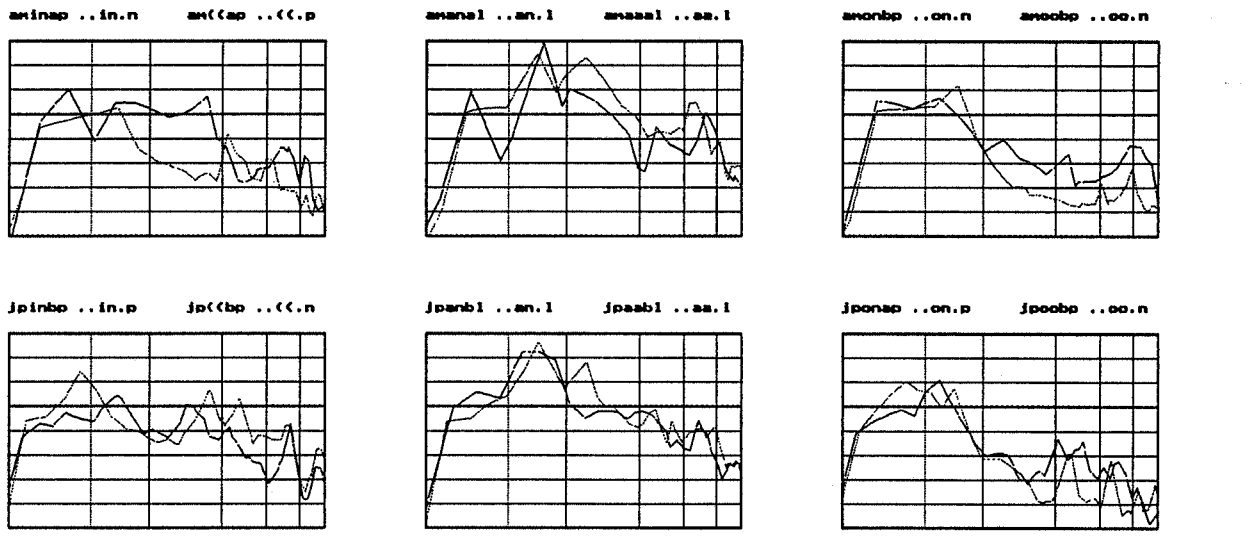


Fig 1 - Quelques exemples de paires nasale (en trait plein) / orale (en pointillé). De gauche à droite : [ɛ̃]/[ɛ], [ɑ̃]/[a], [ɔ̃]/[o]. En haut un locuteur féminin, en bas un locuteur masculin. La force de voix est faible à gauche et à droite, forte au milieu. Les enveloppes spectrales sont représentées en échelle Bark (marques à 0.5, 1, 2, 3, 4 kHz) et en décibels (marques tous les 10dB).

RESULTATS QUALITATIFS

Sur une partie de notre corpus (3 locuteurs et 3 locutrices, 3 forces de voix, deux sessions) nous avons apparié chaque voyelle nasale avec son homologue orale, toutes deux validées perceptivement en ce qui concerne l'identification vocalique. Quelques exemples de ces paires sont représentées sur la fig 1. On rapporte dans ce qui suit ce que les deux éléments de chaque paire ont en commun, ce qui les différencie, en quoi les indices observés confirment ou infirment les résultats publiés à ce jour.

La paire [ɛ̃]-[ɛ]

La caractéristique commune est l'existence d'un creux situé entre 1000 et 1500 Hz, entouré de deux bosses l'une dans la région 400 - 800 Hz (F1 ou F'1), l'autre dans la région 1500 - 2200 Hz (F2 ou F'2).

En ce qui concerne les différences on constate que F1 monte d'environ 150 Hz, F2 descend d'environ 300 Hz, soit pour chaque formant un mouvement de 1 à 2 Bark dans le sens d'une compacité croissante. Le creux qui les sépare est toujours mieux rempli pour la nasale que pour l'orale. Pour certains locuteurs l'effet d'un zéro de nasalité vers 500 Hz est manifeste. Par contre le renforcement de Fn1 à 250 Hz n'apparaît qu'occasionnellement.

Pour transformer un [e] synthétique en son équivalent nasal Hawkins et Stevens mettent un zéro à 500 Hz et un pôle additionnel à 700 Hz. Ceci a pour effet d'augmenter notablement F1 qui passe de 400 à 700 Hz. F2 reste inchangé à 2000 Hz. Même si les fréquences sont différentes (car la voyelle en question est plus proche du [e] français que du [ɛ] qui nous sert de comparaison), le mécanisme semble bien expliquer l'évolution de F1. Mais l'hypothèse selon laquelle F2 n'est pas affecté par la nasalisation semble ici tout-à-fait irréaliste. D'après Lonchamp on passe de [ɛ] à [ɛ̃] en faisant évoluer F1 de 450 à 680 Hz, F2 de 1800 à 1480 Hz, et en ajoutant le premier formant nasal à 280 Hz. Mis à part le dernier point, cette description est en accord avec nos relevés.

La paire [a]-[ɑ̃]

A quelques exceptions près, ces enveloppes des deux voyelles ont en commun un maximum global situé entre 500 et 1200 Hz, et un creux relatif aux alentours de 1,5 à 2 kHz. On ne distingue pas toujours clairement les deux formants du [a]. La partie haute du spectre (au delà de 2 kHz) est souvent renforcée, mais présente une atténuation d'au moins 20 dB par rapport à la région dominante. La partie basse (entre Fo et 500 Hz) est essentiellement plate ou concave, avec une atténuation moindre que la partie haute (entre 5 et 15 dB).

La différence se trouve essentiellement dans l'atténuation de la région 1200 Hz par rapport à la région dominante : cette atténuation est inférieure à 15 dB pour [a], supérieure à 15 dB pour [ɑ̃]. Si l'on prend comme référence l'amplitude du fondamental, l'ensemble du spectre est légèrement atténué pour [ɑ̃] par rapport à [ɑ]. Ceci est plus net en voix faible ou normale qu'en voix forte.

Dans l'ensemble on est assez loin des résultats publiés par Hawkins et Stevens, qui nasalisent le [a] de synthèse en mettant un pôle additionnel à 800 Hz et un zéro à 600 Hz, ce qui a pour effet d'augmenter F1 (qui passe de 700 à 800 Hz, soit environ 1 Bark) tout en laissant F2 inchangé (vers 1200 Hz). Par contre Lonchamp considère que F1 reste inchangé (vers 600 Hz) et que F2 baisse de 1150 à 950 Hz (environ 2 Bark), ce qui correspond mieux à nos observations, avec cependant une assez grande variabilité propre au locuteur et à la force de voix. On retrouve ici une observation de Castelli selon laquelle la résonance Fn1 à 250 Hz apparaît aussi dans les voyelles orales, en particulier dans [a]. Quant au zéro de nasalité vers 500 Hz, il n'est pas observé chez tous les locuteurs.

La paire [ɔ̃]-[o]

Dans tous les cas on observe un maximum global entre Fo et 800 Hz. La partie basse du spectre est en forme de plateau, et la partie haute (au delà de 1000 Hz) est atténuée d'au moins 20 dB pour la voix forte, 25 à 30 dB pour la voix faible.

En ce qui concerne les différences on observe dans la plupart des cas (soit environ 2 fois sur 3) deux maxima à environ 400 et 700 Hz pour [o] ; pour [ɔ̃] on observe un maximum dans la zone 500-600 Hz, et un creux ou un plateau à 400 Hz environ. On trouve des formes dégénérées, en particulier en voix forte, où n'apparaît qu'un seul maximum dans la zone Fo-800 Hz. Dans ce cas on a affaire à [o] si le maximum est au dessous de 500 Hz, et à [ɔ̃] s'il est au-dessus de 600 Hz. Il subsiste quelques cas ambigus. D'une manière générale [ɔ̃] possède plus d'énergie que [o] dans la bande haute.

D'après Hawkins et Stevens on passe de [o] à son équivalent nasal en plaçant un zéro à 500 Hz, ce qui annule le F1 oral défini à 450 Hz, et un pôle à 650 Hz, F2 restant inchangé à 800 Hz. Comme précédemment ceci correspond à nos observations en ce qui concerne F1, mais peut difficilement être confirmé en ce qui concerne F2, très difficile à identifier dans cette région. D'après Lonchamp le F1 oral disparaît et se transforme en formant de nasalité Fn1 à 280 Hz, tandis que F2 passe de 800 Hz pour [o] à 650 Hz pour [ɔ̃]. Ceci correspond bien à nos observations, compte tenu des variations dues au locuteur et à la force de voix.

RESULTATS QUANTITATIFS : SPECTRES DIFFERENTIELS MOYENS

On trouvera fig 2 le "spectre différentiel moyen" (sdm), constitué par la différence entre le spectre de la nasale et celui de l'orale correspondante, calculé sur les 18 séries d'un sous-corpus, soit environ 1/3 de la base de données. Malgré l'aspect un peu sommaire de ce moyennage (aucun appariement manuel ; aucun ajustement des échelles fréquentielles dont on sait qu'elles diffèrent notablement entre voix masculines et voix féminines ; les moyennes peuvent inclure des exemplaires non validés perceptivement), on retrouve dans les sdm certaines des observations rapportées ci-dessus. Pour interpréter les détails du spectre différentiel il faut observer qu'un passage à zéro dans le sens croissant indique un décalage spectral vers l'aigu. On remarquera en particulier la "bosse de compacité" traduisant le rapprochement de F1 et F2 quand on passe de [ɛ] à [ɛ̃], le creux à 1200 Hz caractérisant [ɑ̃] par rapport à [a] et dû à l'abaissement de F2, la paire pôle (à 600 Hz) -

zéro (à 400 Hz) et l'augmentation d'énergie au-dessus de 1 kHz qui caractérisent le passage de [o] vers [ɔ̃].

VALIDATION PAR RECONNAISSANCE AUTOMATIQUE

Pour évaluer la pertinence des connaissances implicitement codées dans les sdm nous avons mené une expérience de reconnaissance automatique. Notre corpus est divisé en 3 sous-corpus a, b et c. Dans chaque sous-corpus on trouve 18 séries de 12 voyelles, provenant de 6 locuteurs (3 hommes et 3 femmes) utilisant les 3 forces de voix. Les locuteurs des corpus a et b sont les mêmes, enregistrés à 6 mois d'intervalle. Les locuteurs du sous-corpus c sont différents mais équilibrés de la même manière. On choisit un sous-corpus d'apprentissage, sur lequel on calcule pour chacune des 3 nasales 3 sdm avec 3 voyelles orales "de base", en l'occurrence [ ], [a] et [o]. On choisit ensuite un sous-corpus de test, qui est examiné série par série. Pour chaque série on constitue un ensemble de 12 spectres de référence, soit les 9 orales de la série, et trois "nasales synthétiques", constituées à partir des 3 orales de base et des sdm appris précédemment. La raison pour utiliser 3 orales de base alors qu'en principe une seule devrait suffire (la plus proche de la nasale considérée), est que même à l'intérieur d'une série le locuteur peut avoir introduit une variabilité notable : en moyennant trois reconstructions de la même nasale à partir de 3 orales différentes on espère compenser dans une certaine mesure ces variations. La reconnaissance proprement dite consiste à comparer chaque nasale "vraie" à l'ensemble des 12 spectres de référence. La distance de deux spectres est calculée par le biais de 12 coefficients cepstraux, selon une technique décrite dans [12] et largement utilisée aujourd'hui dans les systèmes de reconnaissance. Les résultats obtenus en permutant les sous-corpus d'apprentissage et de test sont reportés dans le tableau ci-dessous (valeurs données en pourcentages d'erreur arrondis à l'entier le plus proche, 54 essais par cas):

ref-->	s/corp a	s/corp b	s/corp c
tst --> a	22	26	48
tst --> b	20	15	43
tst --> c	43	44	20

Pour interpréter ces chiffres il faut se souvenir des conditions expérimentales. D'un côté un choix aléatoire (parmi 12 possibilités) mènerait à une valeur de 91%. De l'autre les évaluations perceptives ont montré un taux d'erreur de 15% sur les nasales, au dessous duquel il est peu probable qu'un processus automatique puisse descendre. Les résultats sont donc très bons en auto-cohérence (corpus d'apprentissage identique au corpus de test), encore bons d'une session à l'autre avec les mêmes locuteurs, et nettement moins bons quand on change l'ensemble des locuteurs. Cependant, même dans ce dernier cas, le pourcentage d'erreur moyen qui est de l'ordre de 45% reste très prometteur. D'une part on peut penser qu'environ un tiers de ces erreurs sont dûes à des données mal articulées. D'autre part on a déjà mentionné le fait que la procédure de calcul des sdm est très simple et ne tient pas compte de toutes les informations disponibles dans le corpus d'apprentissage. En tout état de cause ces résultats permettent de valider les observations faites plus haut et illustrées par la fig 2 ci-dessous.

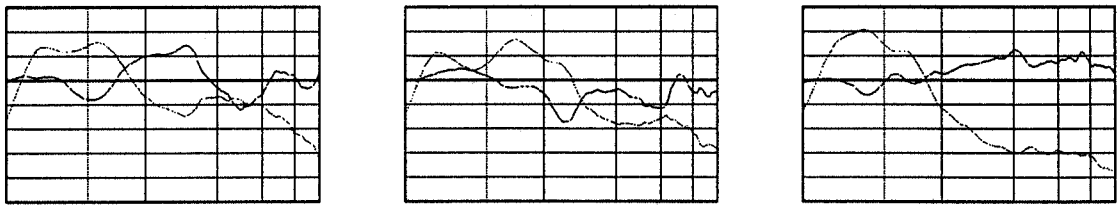


Fig 2 - Spectres différentiels moyens (en trait plein) et voyelles orales moyennes (en pointillé), obtenues sur le sous-corpus b (18 séries : 6 locuteurs et 3 forces de voix). De gauche à droite : sdm [ɛ̃]-[ɛ] et orale moyenne [ɛ], sdm [ɑ̃]-[a] et orale moyenne [a], sdm [ɔ̃]-[o] et orale moyenne [o]. Les sdm sont centrés sur l'ordonnée 50 dB.

CONCLUSION

Cette étude nous a permis de faire le point sur le problème de la nasalité vocalique en français, et sur sa détection par des moyens automatiques dans des conditions peu contraintes (locuteurs non-professionnels, conditions d'élocution relativement libres). Nous nous sommes placés dans la perspective particulière de l'évaluation des nasales par rapport à une ou plusieurs voyelles orales du même locuteur. Les premiers résultats semblent indiquer la possibilité d'identifier les nasales du français dans ces conditions. On notera le fait qu'une étude phonétique de la nasalité fait habituellement appel à une description spectrale en termes de formants ; une telle description est extrêmement difficile à mettre en oeuvre dans un processus automatique, même pour des voyelles orales, dès lors que l'on sort des conditions de laboratoire. C'est pourquoi nous avons préféré rechercher des indices purement spectraux, en l'occurrence les différences spectrales caractérisant chaque nasale par rapport à l'orale correspondante. Dans cette approche le passage du spectre brut à une enveloppe phonétiquement significative, indépendante du fondamental et du bruit de fond, constitue une phase critique du travail d'analyse. L'hypothèse générale de notre travail, qui consiste à ne considérer que le spectre de stabilité, est évidemment provisoire dans la mesure où il est établi que la reconnaissance d'une voyelle est aussi liée à son évolution temporelle [13].

Nous n'avons pas manqué, tout au long de cette étude, de nous poser la question de l'existence d'indices absolus de la nasalité vocalique, indépendants de la voyelle nasale considérée. Il est sans doute trop tôt pour pouvoir répondre à cette question. Nos relevés, confrontés aux connaissances publiées, montrent notamment que les répercussions spectrales de la paire pôle-zéro dans la région de F1 contribuent notablement à distinguer la voyelle nasale de sa correspondante orale. Ceci corrobore partiellement les expériences de Hawkins et Stevens. Pourtant il semble bien que l'hypothèse d'invariance de F2 ne soit pas vérifiée dans notre corpus, en particulier pour les paires [ ]-[a] et [ ]-[ ], ce qui rejoint certaines des conclusions de Lonchamp. On peut douter de la valeur universelle, du point de vue perceptif, de Fn1, de la faible valeur ou de l'élargissement de F1, de l'amortissement général du spectre, voire même de la perception d'un zéro en tant que tel dans la région 400-600 Hz. Par contre pour le français il semble que la perception des voyelles nasales ne soit pas fondamentalement différente de celle des voyelles orales : nous rejoignons Lonchamp dans l'idée que les unes comme les autres peuvent trouver leur place spécifique dans une classification bidimensionnelle, dont les dimensions perceptives, qui restent à préciser, ne sont sans doute pas exactement identiques aux formants pris en considération par la théorie de la production.

REFERENCES

1 - Maeda S. : "Acoustics of vowel nasalization and articulatory shifts in French nasal vowels", in Phonetics and Phonology, vol 5, "Nasals, Nasalization, and the Velum", eds M.K.Huffman and R.A.Krakow, Ac. Press Inc., 1993.

2 - Castelli E. : "Caractérisation acoustique des voyelles nasales du français", thèse Inst. Nat. Polyt. de Grenoble, juin 1989.

3 - Glass J.R. and Zue V.W : "Detection of nasalized vowels in American English", IEEE ICASSP, Tampa,1985.

4 - Lonchamp F. : "Recherche sur les indices perceptifs des voyelles orales et nasales : application à la structure du système vocalique français et de diverses autres langues", thèse 3e cycle, univ. Nancy II, mai 1978.

5 - Delattre P. : "Les attributs acoustiques de la nasalité vocalique et consonantique", Studia Linguistica 8, 103-109, 1954.

6 - Hawkins S. and Stevens K.N. : "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels", J. Acoust. Soc. Am 77(4), April 1985.

7 - Liénard J.S. et Di Benedetto, M.G. : "Evaluation perceptive d'un corpus de voyelles françaises émises isolément par plusieurs locuteurs selon diverses forces de voix", 19e Journées d'Etude sur la Parole de la SFA, Bruxelles, mai 1992.

8 - Serra X. : "A system for sound analysis, transformation and synthesis based on a deterministic plus stochastic decomposition", Ph.D. dissertation, Stanford University, Oct. 1989.

9 - McAulay R.J. and Quatieri T.F. : "Speech analysis/synthesis based on a sinusoidal representation", IEEE Tr. on ASSP, vol ASSP 34, 744-754, 1986.

10- Grau-Grovel S. : "Utilisation de la représentation par formes d'ondes formantiques et de la représentation sinusoïdale dans l'élaboration d'un synthétiseur paramétrique de parole", thèse, univ. Paris XI, nov. 1992.

11 - Schroeder M.R., Atal B.S. and Hall J.L. : "Optimizing digital speech coders by exploiting masking properties of the human ear", J. Acoust. Soc. Am 66(6), Dec. 1979.

12 - Davis S.B and Mermelstein P. : "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Tr. on Acoustics, Speech and Signal Processing, vol ASSP-28 n° 4, Aug. 1980.

13 - Di Benedetto M.G. : "Vowel representation : some observations on temporal and spectral properties of the first formant frequency", J.Acoust.Soc.Am. 86 (1), July 1989.

MATRICES PHONETIQUES ET MATRICES PHONOLOGIQUES ARABES

B. BOUDRAA*, S.A. SELOUANI*, M. BOUDRAA*, B. GUERIN**

* LCP, Institut d'électronique, USTHB, BP.32, El-Alia, Alger, Algérie.
** ICP, U.A. CNRS N° 368, INPG/ENSERG & Université Stendhal,
46 Avenue Félix Viallet, 38031, Grenoble cedex, France.

Résumé

Nous appliquons ici la méthode de Boë et Tubach pour l'extraction des traits distinctifs de la langue arabe à partir d'une description phonétique de cette langue. A cet effet, nous prenons en compte la fréquence d'occurrence des sons de l'arabe donnée par Mrayati et Moussa pour établir la hiérarchisation entre les traits. Nous utilisons ensuite les règles de Boë et Tubach pour passer à la matrice phonologique de l'arabe et pour retourner de celle-ci à la matrice phonétique.

Mots-clés: Matrice phonétique, matrice phonologique, fréquene d'occurrence, entropie.

INTRODUCTION

Les systèmes de reconnaissance "basées connaissances", les tests d'intelligibilité de la parole de synthèse, etc., s'appuient sur le savoir acquis en sciences phonétiques. Deux modèles ont grandement influencé le passé dans ce domaine: l'école de Prague (Jakobson) et l'école américaine (Chomsky et Halle). Pour les deux écoles, la description des sons d'une langue passe par la définition de traits acoustiques pertinents. Jakobson classe les sons à l'aide de 12 traits binaires distinctifs minimaux et dont la validité serait universelle. Les 12 oppositions possibles sont décrites en termes acoustiques et articulatoires. Chomsky et Halle (1968) ayant trouvé des contradictions dans les traits de Jakobson ont défini d'autres traits purement articulatoires et ne les ont pas qualifiés d'universels. Ce système est redondant et les traits ne sont pas binaires. En fait, le problème de la hiérarchisation des traits s'est toujours posé. A cet effet, une approche interessante a été effectuée par Boë et Tubach (1986). Cette hiérarchie se rapproche des travaux de Van Den Broecke (1976) et a eu pour champ d'application la taxinomie établie par Rossi (1983) pour le français. Cette taxinomie avait l'avantage d'avoir été testée dans des systèmes de reconnaissance automatique de la parole (Caelen,1979; Fonsale, 1984). Dans leur approche, Boë et Tubach effectuent une dichotomie entre les voyelles et les consonnes qui permet entre autres de faciliter la mise en oeuvre de leur procédure.

En ce qui concerne la langue arabe et à des fins d'application en reconnaissance automatique et en évaluations subjectives en synthèse, nous présenterons ici une taxinomie séparant les consonnes et les voyelles. Nous effectuons ensuite une hiérarchisation des traits en se basant sur l'équilibre phonétique de la langue arabe effectué par Mrayati (1987) pour les consonnes et Moussa (1987) pour les voyelles. Nous reprenons enfin la méthode de Boë et Tubach pour passer de la matrice phonétique arabe à celle phonologique et inversement.

MATRICE PHONETIQUE DE L'ARABE

Que sait-on de la réalisation phonétique de la langue arabe? D'abord, il faut distinguer les travaux portant sur l'arabe litteraire classique de ceux effectués sur l'arabe dit moderne. Les travaux de Cantineau (1941 et 1946) peuvent constituer un point de départ solide pour toute réflexion, mais il faut tenir compte que ceux-ci sont une description de l'arabe classique effectuée à partir du Saint Coran et d'informations fournies par d'anciens grammairiens arabes tel Sibawayhi. Les travaux de Roman (1975) peuvent aussi s'inscrire dans cette ligne, mais, ce dernier a eu le mérite de conclure ses travaux sur l'arabe classique, par une approche moderne tout en évoquant l'évolution du système. C'est ainsi que Roman (1975,p.216) note l'existence de 10 traits acoustiques du modèle élaboré par Jakobson. Il complète ces traits par deux autres qu'il nomme respectivement complexe ~ simple et soufflé ~ non soufflé. Pour l'arabe moderne, il faut également noter les travaux normatifs d' El-Ani (1970) et ceux de Ghazeli (1979). Les deux auteurs obtiennent des résultats appuyés sur des méthodes expérimentales radiocinématographiques et sonagraphiques. El-Ani considère des locuteurs d'Iraq. Ghazeli considère 12 locuteurs masculins originaires de six pays arabes (Algérie, Tunisie, Libye, Egypte, Jordanie, Iraq). De ce dernier travail, nous avons retenu la notion de tension et de laxité introduite dans le cas des voyelles arabes et qui se trouve en correspondance avec celle que Jakobson a déja cité dans ses *Preliminaries,(1963, p.36)*. Aussi, notre matrice contiendra ce trait, dans le cas des consonnes et des

voyelles. Nous proposons ainsi, les matrices phonétiques suivantes (Tableaux 1 et 2).

	AG	IT	CT	DF	TN	VC	ST	NZ	BM
n ن	+	+	+	-	-	+	-	+	-
l ل	+	+	+	-	-	+	-	-	-
r ر	+	+	-	+	-	+	-	-	-
ʒ ج	+	+	-	+	-	-	+	-	-
$t_{\bullet}$ ط	+	+	-	-	+	-	-	-	+
t ت	+	+	-	-	+	-	-	-	-
$d_{\bullet}$ ض	+	+	-	-	-	-	-	-	+
d د	+	+	-	-	-	-	-	-	-
ʃ ش	+	-	+	+	+	-	+	-	-
z ز	+	-	+	+	-	+	-	-	-
$s_{\bullet}$ ص	+	-	+	-	+	-	+	-	+
s س	+	-	+	-	+	-	+	-	-
θ ث	+	-	+	-	+	-	-	-	-
z ز	+	-	+	-	-	-	+	-	-
$ð_{\bullet}$ ظ	+	-	+	-	-	-	-	-	+
ð ذ	+	-	+	-	-	-	-	-	-
m م	-	+	+	-	-	+	-	+	-
k ك	-	+	-	+	+	-	-	-	-
q ق	-	+	-	+	-	-	-	-	-
ʔ ء	-	+	-	-	+	-	-	-	-
b ب	-	+	-	-	-	-	-	-	-
x خ	-	-	+	+	+	-	+	-	-
ħ ح	-	-	+	+	+	-	-	-	-
ɣ غ	-	-	+	+	-	-	+	-	-
ε ع	-	-	+	+	-	-	-	-	-
f ف	-	-	+	-	+	-	+	-	-
h ه	-	-	+	-	+	-	-	-	-
w و	-	-	+	-	-	+	-	-	-

Tableau 1: Matrice phonétique des consonnes de la langue arabe, à partir des traits AIGU, INTERROMPU, CONTINU, DIFFUS, TENDU, VOCALIQUE, STRIDENT, NASAL et BEMOLISE.

	/a/	/a:/	/i/	/i:/	/u/	/u:/
DF	-	-	+	+	+	+
TN	-	+	-	+	-	+
AG	-	-	+	+	-	-

Tableau 2: Matrice phonétique des voyelles de la langue arabe à partir des traits DIFFUS, TENDU et AIGU.

Trois traits décrivent les voyelles et neuf traits classent les consonnes. En fait, chaque trait n'a pas la même entropie et, par conséquent, n'apporte pas la même quantité d'information. Par ailleurs, dans un système à choix binaire, l'entropie s'écrit:

$$H = -p_- \log_2 p_- - p_+ \log_2 p_+$$

$p_+ et p_-$ sont respectivement les sommes des probabilités des réponses positives et négatives pour chaque trait. L'entropie est ici maximale si $p_+ = p_- = 1/2$ et minimale si $p_+=1$ et $p_-=0$ ou inversement.

Le calcul de cette entropie, prenant en compte la fréquence d'occurrence des sons de la langue arabe, permettra d'établir une hiérarchie entre les traits.

OCCURRENCES DES SONS DE L'ARABE

Dans ses études détaillées sur les racines et les mots de la langue arabe et effectuées sur trois dictionnaires arabes: *Al-Sehah , Lissan Al-ɛarab et Taə Al- ɛarous,* A.H. Moussa donne des fréquences d'occurrences des voyelles longues et brèves de la langue arabe ainsi que les fréquences des consonnes. Pour les voyelles, il ajoute la féquence du diacritique "pause" (Soukoun).

Des études analogues effectuées par Mrayati et portant sur le dictionnaire *Lissan Al-ɛarab* utilisé par Moussa et sur quatre autres dictionnaires: *ɜamharatu Al-luɣah, tahdibu Al-luɣah, Al-muħkam et al-qamus* $Al-muħi_{\bullet}t$ ont conduit à des données différentes pour les consonnes. Cependant, Mrayati ne donne pas de résultats pour les voyelles. Aussi, nous avons retenu les fréquences d'occurrence des voyelles longues et brèves données par Moussa et correspondant à 73.709 % (26.291 % ont été affectées à la pause). Nous avons délibérément omis la pause (car ne répondant pas à la définition classique des voyelles) et déduit les fréquences des autres voyelles. Le tableau 3 donne les fréquences obtenues.

Voyelle	M-fréq. (%)	B-fréq. (%)
Fetħa /a/	32.473	44.056
Dhammah /u/	9.487	12.871
Kasrah /i/	13.095	17.766
Alif al-mad /a:/	11.352	15.401
Waw al-mad /u:/	4.099	5.561
Ya: al-mad /i:/	3.203	4.345
total	73.709	100

Tableau 3: Fréquences d'occurrences des voyelles. B-fréq(%) sont celles M-fréq de Moussa, ramenées en % après omission de la pause.

En ce qui concerne les consonnes, nous avons considéré les fréquences établies par Mrayati car, d'une part, elles ont été obtenues sur un plus grand nombre de dictionnaires et, d'autre part, ces dictionnaires sont plus récents. Le tableau 4 donne les fréquences des consonnes retenues.

Consonnes	Fréquences (%)
n ن	6.11
l ل	6.12
r ر	7.90
ʒ ج	3.93
$\dot{t}$ ط	3.01
t ت	2.34
$\dot{d}$ ض	1.49
d د	4.43
ʃ ش	3.22
j ي	2.40
$\dot{s}$ ص	2.18
s س	4.17
θ ث	1.97
z ز	2.76
$\dot{\partial}$ ظ	0.61
∂ ذ	1.52
m م	5.51
k ك	2.96
q ق	4.58
? ء	2.25
b ب	6.02
x خ	2.81
ħ ح	3.63
γ غ	1.92
ε ع	5.02
f ف	4.10
h ه	3.43
w و	3.61

Tableau 4: Fréquence d'occurrence des consonnes.

Le tableau 5 donne les entropies obtenues par trait, respectivement pour les voyelles et les consonnes. Le classement des traits respectera par la suite les valeurs des entropies obtenues.

Traits	V-Entropie	C-Entropie
AG	0.762	0.996
IT	-	0.987
CT	-	0.964
DF	0.934	0.961
TN	0.816	0.943
VC	-	0.901
ST	-	0.813
NZ	-	0.518
BM	-	0.377

Tableau 5: Entropies par trait pour les voyelles et pour les consonnes.

MATRICE PHONOLOGIQUE.

Afin d'obtenir la matrice phonologique, nous opérons de la même manière que Van Den Broecke pour la première étape et Boë et Tubach pour la suite. Van Den Broecke (1976) tient compte d'un facteur de pondération lié au nombre d'occurrence d'un trait dans les possibles jeux nécessaires et suffisants pour caractériser un son. La démarche de Boë et Tubach consiste à reprendre les oppositions une à une, à vérifier à l'aide de la matrice phonétique si cette première spécification est suffisante et, dans le cas contraire, ajouter une spécification de plus, en respectant la hiérarchie des traits distinctifs. Ainsi, On chiffre d'abord pour chaque classe la "distance" phonétique en effectuant une comparaison deux à deux de tous les traits dont les sons diffèrent.

On note que dans 7 cas, un seul trait permet de faire la distinction entre les voyelles. Il s'agit des traits donnés dans le tableau 6 suivant:

Trait	V-Opposition
DF	u / a u: / a:
TN	a: / a i: / i u: / u
AG	i / u i: / u:

Tableau 6: Oppositions binaires des voyelles de l'arabe.

Pour les consonnes, on distingue les 29 oppositions données dans le tableau 7 suivant:

Trait	C-Opposition
AG	n / m ʃ / x t / ? d / b s / f θ / h
DF	k / ? ʃ / s x / f $\hbar$ / h q / b
TN	k / q x / γ $\hbar$ / ε t / d $\dot{t}$ / $\dot{d}$? / b s / z θ / ∂
ST	γ / ε x / $\hbar$ s / θ z / ∂ f / h
NZ	n / l
BM	$\dot{t}$ / t $\dot{d}$ / d $\dot{s}$ / s $\dot{\partial}$ / ∂

Tableau 7: Oppositions binaires des consonnes de l'arabe.

Le tableau 8 donne les distances entre toutes les voyelles et toutes les consonnes de la langue arabe. Par la suite, les matrices phonologiques se remplissent d'abord par ces traits nécessaires. Puis, nous utilisons la démarche de Boë et Tubach pour compléter la matrice phonologique par les traits "suffisants" en respectant l'ordre hiérarchique établi à l'aide de l'entropie.

	h	ʔ	f	ð	ð̣	θ	z	s	ṣ	b	ḍ	d	ṭ	t	ɛ	ħ	ɣ	x	ʃ	q	k	ə	w	j	r	l	m
n	5	5	6	3	4	4	3	5	6	4	4	3	5	4	5	6	6	7	6	5	6	5	3	3	3	1	1
m	4	4	5	4	5	5	4	6	7	3	5	4	6	5	4	5	5	6	7	4	5	6	2	4	4	2	
l	4	4	5	2	3	3	3	4	5	3	3	2	4	3	4	5	5	6	5	4	5	4	2	2	2		
r	6	4	7	5	5	5	5	6	7	3	3	2	4	3	4	5	5	6	5	2	3	2	4	2			
j	4	6	5	2	3	3	3	4	5	5	5	4	6	5	2	3	3	4	3	4	5	4	2				
w	2	4	3	2	3	3	3	4	5	3	5	4	6	5	2	3	3	4	5	4	5	6					
ə	6	4	5	4	5	5	3	4	5	3	3	2	4	3	4	5	3	4	3	2	3						
k	3	1	4	5	6	4	6	5	6	2	4	3	3	2	3	2	4	3	4	1							
q	4	2	5	4	5	5	5	6	7	1	3	2	4	3	2	3	3	4	5								
ʃ	3	5	2	3	4	2	2	1	2	6	6	5	5	4	3	2	2	1									
x	2	4	1	4	5	3	3	2	3	5	7	6	6	5	2	1	1										
ɣ	3	5	2	3	4	4	2	3	4	4	6	5	7	6	1	2											
ħ	1	3	2	3	4	2	4	3	4	4	6	5	6	4	1												
ɛ	2	4	3	2	3	3	3	4	5	3	5	4	6	5													
t	3	1	4	3	4	2	4	3	4	2	2	1	1														
ṭ	4	2	5	4	3	3	5	4	3	3	1	2															
d	4	2	5	2	3	3	3	4	5	1	1																
ḍ	5	3	6	3	2	4	4	5	4	2																	
b	3	1	4	3	4	4	4	5	6																		
ṣ	3	5	2	3	2	2	2	1																			
s	2	4	1	2	3	1	1																				
z	3	5	2	1	2	2																					
θ	1	3	2	1	2																						
ð̣	3	5	4	1																							
ð	2	4	3																								
f	1	3																									
ʔ	2																										

a)

b)

				u:
			1	u
		2	1	i:
	1	1	2	i
3	2	2	1	a:
1	2	3	1	2	a
a:	i	i:	u	u:	

Tableau 8: Distances entre a)- les consonnes b)-les voyelles

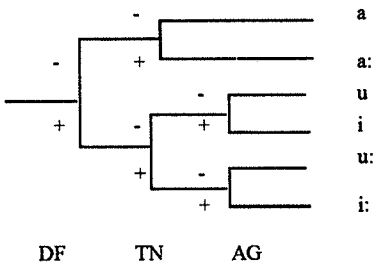
Dans le cas des consonnes, il fallait rajouter les traits donnés dans le tableau 9. Pour les voyelles, les traits distinctifs nécessaires étaient également suffisants pour établir la matrice phonologique (tableau 10) ainsi que l'arbre I. Dans ce dernier, la branche TN a précédé celle AG à cause des valeurs de leurs entropies. Ceci a conduit en fin de branche des oppositions u / i et u: / i: au lieu des u / u: et i / i: habituelles.

Trait	C-Opposition complémentaire
AG	l / w r / q ð / ɛ z / ɣ
IT	r / j t / θ d / ð ḍ / ð̣ m / w k / ħ ʔ / h q / ɛ
CT	l / r m / b
DF	j / ð r / d ɛ / w
TN	ʃ / j h / w s / ð̣
VC	r / ə

Tableau 9: Oppositions complémentaires des consonnes

	/ a /	/ a: /	/ i /	/ i: /	/ u /	/ u: /
DF	-	-			+	+
AG			+	+	-	-
TN	-	+	-	+	-	+

Tableau 10: Matrice phonologique pour les voyelles.

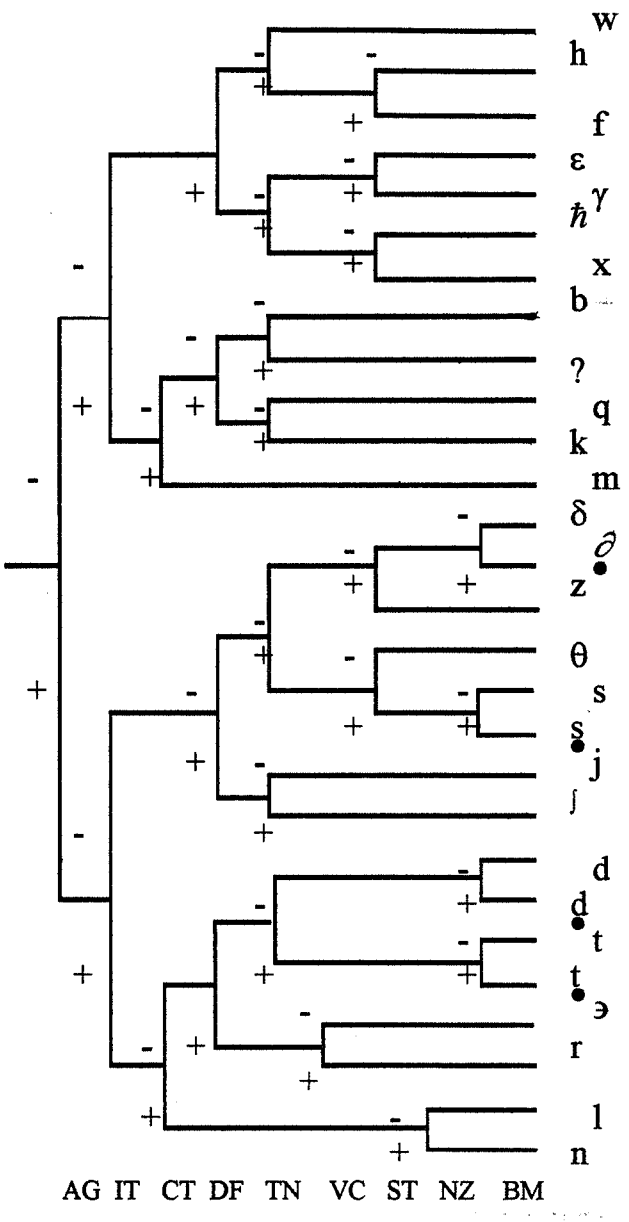


Arbre I: Système Vocalique Arabe

Finalement, nous obtenons la matrice phonologique de la langue arabe donnée par le tableau 11 suivant:

	AG	IT	CT	DF	TN	VC	ST	NZ	BM
n ن	+							+	
l ل	+	+	+					-	
r ر	+		-	+		+			
ʒ ج						-			
$t_{\bullet}$ ط					+				+
t ت	+	+			+				-
$d_{\bullet}$ ض		+			-				+
d د	+	+		-	-				-
ʃ ش	+			+	+				
j ي		-		+	-				
$s_{\bullet}$ ص					+				+
s س	+			-	+		+		-
θ ث	+	-			+		-		
z ز	+				-		+		
$ð_{\bullet}$ ظ		-			-				+
ð ذ	+	-		-	-		-		-
m م	-	+	+						
k ك		+		+	+				
q ق	-	+		+	-				
? ء	-	+		-	+				
b ب	-		-	-	-				
x خ	-			+	+		+		
ħ ح		-		+	+		-		
γ غ	-				-		+		
ε ع	-	-		+	-		-		
f ف	-			-			+		
h ه	-	-		-	+		-		
w و	-	-		-	-				

Tableau 11: Matrice phonologique pour les consonnes



Arbre II: Système consonantique

Nous avons exploité par la suite la règle d'implication établi par Boë et Tubach pour établir le passage inverse, c'est à dire de la matrice phonologique à la matrice phonétique.

CONCLUSION.

Tout en mettant en évidence une hiérarchie entre les traits dans le cas de la langue arabe, ce travail met au point une démarche que nous allons exploiter dans le cadre de la reconnaissance en langue arabe. Par ailleurs, les oppositions obtenues vont nous servir à mettre au point des matériaux liguistiques de tests subjectifs en synthèse et en codage dans le cas de la langue arabe.

BIBLIOGRAPHIE.

S.H. AL-ANI (1970), "Arabic phonology: An Acoustical and Physiological Investigation", Mouton, The Hague, 104p.

L.J. BOE, J.P. TUBACH (1986), "Des matrices phonétiques aux matrices phonologiques et vice versa", Bulletin de l'Institut de phonétique de Grenoble, Vol. 15, pp. 135-155.

J. CAELEN (1979), "Un modèle d'oreille, analyse de la parole continue, reconnaissance phonémique", Thèse de doct. d'Etat. Sciences, Toulouse.

J. CANTINEAU (1946), "Esquisse d'une phonologie arabe classique", BSL, N°126, pp.93-140.

J. CANTINEAU (1960), " Cours de phonétique arabe", Klincksieck, (1ère ed. 1941). Paris.

N. CHOMESKY, M. HALLE (1968), "The sound Pattern of English", Harper and Row, New York.

P. FONSALE (1984), "Traits multilocuteurs et leur utilisation dans un système de reconnaissance de mots isolés", 11ème ICA.

S. GHAZELI (1979), "Du statut des voyelles en Arabe", Analyses-théories, Etudes arabes, N° 2-3, 1979, pp.199-219.

R. JAKOBSON, C.G.M. FANT, M. HALLE (1963), "Preliminaries to speech analysis: The distinctive featues and their correlates", Cambridge, MIT Press.

A.H. MOUSSA (1987), "Computer applications to Arabic Roots and Arabic words", Applied Arabic Linguistics and Signal and Information Processing, Hamshire Publishing.

M. MRAYATI (1987), " Statistical Studies of Arabic Roots ", Applied Arabic Linguistics and Signal and Information Processing, Hamshire Publishing.

A. ROMAN (1975), "Le système phonologique de l'arabe classique au VIIIème siècle d'après le Kita:b de Si:bawayhi", Travaux de l'Institut de phonétique d'Aix", Vol.2, pp.203-232.

M. ROSSI (1983), "Niveaux de l'analyse phonétique: nature et structuration des indices et des traits", Speech communication, 2, pp. 91-106.

M.P.R. VAN DEN BROECKE (1976), "Hierarchies and rank orders in distinctive features", Van Gorcum Assen, Amsterdam.

ORALITÉ, CONTEXTE ET COGNITION SUR LE TERRAIN CRÉOLE RÉUNIONNAIS

GILLETTE STAUDACHER-VALLIAMEE

L.A.C.I.T.O./C.N.R.S. Paris

Résumé

Cette contribution présente des données collectées dans le cadre d'une enquête prosodique organisée en échantillonage représentatif et menée auprès de 20 locuteurs créolophones natifs de La Réunion, département insulaire francophone de L'Océan Indien.
L'outil utilisé repose sur l'élaboration d'une fiche de prononciation construite autour de 126 items extraits du fonds lexical commun constitutif de ce parler créole. Ils ont été choisis selon des critères phonologiques (place des phonèmes vocaliques orals et nasals dans le mot), syllabiques (items mono-, di-, polysyllabiques, composés figés), sémantiques (vocabulaire domestique, culinaire, médical ancré dans la réalité socio-historique) de la culture insulaire.
L'hypothèse sous-jacente à cette expérimentation était de vérifier la nature du lien entre habitudes articulatoires (phonèmes) marquant l'oralité en contexte de système (phonologique) et l'écriture ou la représentation graphique des pertinences de l'oralité.
Les premiers résultats observés soulignent le rôle déclencheur et moteur des schèmes de l'oralité dans le décodage nécessaire à ce type de perception et de production. Activité rarement pratiquée dans la communauté créolophone, la lecture à haute voix en créole réunionnais, présuppose une compréhension pluridimensionnelle et diversifiée conditionnée par plusieurs catégories de connaissances: celles que doit avoir le sujet créolophone natif pour appréhender le contexte socio-culturel véhiculé par les mots choisis. Le savoir et savoir-faire liés à l'origine de la tradition orale (française, malgache, indo-portugaise) et sa représentation écrite sont de nature plus technique.

Mots-clés: créole réunionnais, enquête prosodique, phonologie, oralité, contexte, cognition.

INTRODUCTION

L'étude du lien entre le son et la lettre revêt un intérêt tout particulier quand elle s'applique aux différents créoles (Valdman, A., 1977). Les travaux de synthèse réalisés à ce jour rappellent que l'un des dénominateurs communs à la constitution de ces idiomes est bien la rencontre, dans l'histoire et la géographie des déplacements humains, d'au moins deux langues différentes. Les traits constitutifs de cet idiome permettent de penser que les divers apports linguistiques ont laissé dans le système général de la langue des traces d'habitudes articulatoires et crée des contraintes engendrant un syncrétisme à plusieurs niveaux (Staudacher-Valliamee, G., 1993).

On peut comprendre qu'une description phonologique fournisse les éléments de base à une réflexion sur la perception et la production de la parole chez les locuteurs dont le parler créole est marqué par l'histoire des langues du peuplement. N'oublions pas que le créole est la langue première d'une population créolophone native regroupant environ 599 000 habitants. Après avoir examiné ce qui constituait en contexte de synchronie dynamique l'unité et la diversité des usages phonologiques à la Réunion (Staudacher-Valliamee, G., 1990, 1991a, b, 1992a, 1993), l'auteur a élaboré une fiche d'items représentatifs qu'elle a soumis à des informateurs créolophones pour qu'ils les prononcent ou les lisent à haute voix.
L'activité de lecture met le locuteur en situation de décodage alors que les actes d'écriture mobilisent l'encodage. La prise en compte des unités linguistiques minimales en contexte primaire (place du phonème dans le mot discriminé selon sa structure phonologique, syllabique, syntaxique et sémantique) peut apporter des informations utiles aux processus qui interagissent lors du déroulement de cette activité.
Cette contribution exposera les principes linguistiques (phonologiques, syntaxiques et didactiques) qui ont étayé l'élaboration de la fiche en tant qu'instrument d'investigation. Elle résume les premiers résultats déclenchés par la stimulation scripturale pendant la lecture à haute voix. Ils confèrent à l'enquête une valeur expérimentale. En effet, l'échantillonnage montre quels mécanismes semblent en œuvre chez le lecteur créolophone réunionnais (lien entre mémoire auditive et projection scripturale). Il livre des indices nécessaires à une enquête devant être contrôlée en laboratoire dans une étape ultérieure, ce qui approfondirait l'analyse du lien entre oralité, contexte et cognition dans le situation de contact de langues définitoire de ce créole.

L'ORALITE EN CONTEXTE DE SYSTEME PHONOLOGIQUE COMMUN

Sachant que le créole n'est pas enseigné en milieu scolaire et qu'aucune pratique écrite ne sous-tend officiellement sa généralisation dans l'espace insulaire, il est encore indispensable d'appliquer des méthodes d'une linguistique de terrain. pour saisir les traces de l'activité langagière véhiculée par le tradition orale.
L'analyse phonologique de 18 idiolectes de 90 mns soit 27 heures de corpus, avait permis d'identifier un système phonologique commun (Staudacher-Valliamee, G., 1989a, 1990, 1991a, 1992b). Par système phonologique commun ou moyen, il faut entendre celui des phonèmes consonantiques.

L'apport de cette approche empirico-déductive est double. Elle met en relief la structure d'un système conditionné par des facteurs internes et externes. Elle éclaire le rôle déterminant du contexte que l'oralité définit et fait intervenir à plusieurs niveaux. L'analyse des habitudes articulatoires d'un créole est menée d'abord en terme de hiérarchie des pertinences, intégrant la place du phonème dans le mot, sa fréquence (dans le lexique et le discours), son degré de stabilité et le rendement fonctionnel des oppositions. L'épreuve de commutation constitue en soi un premier contrôle théorique de la part du phonologue.

Pour mesurer l'espace situé entre la notation phonologique et la représentation graphique, on pouvait partir du nœud rattachant le système commun (l'unité) à la dynamique de la diversité.

La description avait identifié pour le réunionnais la centralité vocalique (/i/ ~ /ï/, /e/ ~ /ë/), des consonnes apico-alvéolaires (/s/ ~ /ṡ/). La place de ces pertinences dans la dynamique générale du système insulaire est instructive, car on voit que ces traits situés au centre, pourraient servir d'alternative aux pertinences situées aux extrémités du système (ex: la consonne sourde /s:/, les chuintantes simples et complexes. Le concept de système commun ou moyen peut donc approfondir la recherche du noyau alphabétique potentiel en aidant à circonscrire les données contextuelles plus larges, mais non moins déterminantes dans le cas de ce créole. Ce dernier atteste en effet plusieurs pratiques écrites relativement récentes dont aucune n'est à ce jour officialisée. Par ailleurs, le dépouillement des fiches biographiques mettait en évidence un trait commun à tous les informateurs, à savoir la pratique orale du créole dès la naissance (Staudacher-Valliamee, G., 1989b, 1992b). Il apparaît que malgré l'absence de codification écrite officielle, le créole en tant que langue maternelle ou première conditionne la relation des lecteurs et des scripteurs créolophones natifs à tout code écrit. Il peut s'agir du français, langue officielle dans le Département d'Outre-Mer, du créole et des composantes linguistiques définitoires dès l'origine de son fonctionnement (français du 17ème siècle, malgache, indo-portugais). C'est la réalisation (écrite et orale) d'une fiche de prononciation qui dans l'enquête indique le plus clairement la nature des mécanismes et des conditionnements suceptibles d'interagir dans les processus de décodage.

LA REPRESENTATION SCRIPTURALE DU CONTEXTE ORAL

L'élaboration d'une fiche de prononciation (ou prosodique) a été envisagée pour voir si certaines pertinences vocaliques comme la nasalité (ɐ̃, ã, õ), la longueur et la postériorité (a, o, u) constituaient linguistiquement un accent réunionnais. Nous parlons ici de prosodie, car la description phonétique et phonologique des données collectées avait montré que la répartition de la quantité vocalique à l'intérieur du mot engendrait un effet d'accent quantitatif de durée, alors qu'à l'initiale et à la finale, on percevait un accent d'énergie (Staudacher-Valliamee, 1991 b). Aussi, dans les 126 items choisis pour leur degré maximal de représentativité, c'est la place de la voyelle qui a servi de premier critère déterminant. Le contexte prosodique réunit ici les réalités supra-segmentales inhérentes à la chaîne parlée et constitutives de la forme des mots créoles.

Les autres discriminants ont été dans l'ordre décroissant: la combinaison des phonèmes, la structure syllabique et l'ordre fixe des unités lexicales pour le sens de la détermination (mots simples, composés, figés), leur appartenance au fonds lexico- culturel commun.

On remarque tout de suite que ce type d'outil mobilise d'abord le noyau des pertinences et des structures du système linguistique concerné sans l'extraire de la culture et de l'histoire qui l'enracinent dans le contexte insulaire. L'hypothèse était simple: comme chacun parle avant d'écrire, il fallait cerner le rôle des habitudes articulatoires dans le passage à la lecture à haute voix, se rattachant elle-même à la visualisation des sons.

Dans le traitement linguistique du mot, c'est le critère du fonds lexical commun qui a été retenu. Les mots correspondaient à l'expérience quotidienne vécue par toute la communauté créolophone (ex: japana "plante médicinale", longani "longane", brèdd morèll "plante, légume"). Pour le vocabulaire culinaire, on peut citer tri "nettoyer", kwi "cuit", boui "bouilli", ainsi que des viande et poisson (ex: volay "poulet", zouritt "poulpe") ou alors des ustansiles (ex: lavane "van", pilon "mortier").

Le critère sémantique est inséparable du contexte au sens large de connaissances (Baddeley, 1993). C'est surtout pour leur valeur de référents culturels que ces mots déjà activés en mémoire quotidienne ont été rappelés en mémoire scripturale. Il est bien établi que tout créole se caractérise par l'existence, dans son lexique, d'un vocabulaire fortement lié au contexte sociohistorique. Ces items sont particulièrement utiles dans les enquêtes linguistiques, car l'enquêteur peut compter sur leur impact psycho-affectif important pour mener à bien un test linguistique voulant éviter les sujets intimes susceptibles de gêner les informateurs. Ils interpellent également toutes les générations, aussi bien les hommes que les femmes.

Le critère phonologique reste le plus important quand on en arrive à l'étude du lien entre le son et la lettre, entre la prononciation et l'écriture, c'est-à-dire aux principes de la représentation. Le premier volet de la grille rassemblait par exemple 40 mots en /i/ parce que la fréquence de cette voyelle est maximale dans le créole réunionnais. Le deuxième volet présentait des mots contenant des voyelles postérieures /u, o, a/ et des nasales /ɐ̃, õ, ã/. Selon le critère syllabique, il importait aussi de distinguer entre les monosyllabiques (fritt "frit", trimm "trimer", soukk "attraper"), les dissyllabiques (lanti "lentille", mizèr "pauvre"), les polysyllabiques (kankrola "cafard", landòrmi "caméléon", lankostikk "encostique"). A la suite des mots simples apparaissaient les synthèmes (ex: kari volay, rougay sosiss, bouyon brèdd morèll.

ACTIVITE LANGAGIERE ET COGNITIVE DIFFERENCIEE

Pour éviter que des informateurs soient tout de suite exclus à cause de l'illettrisme ou de l'absence de scolarisation en français, la présentation matérielle de la fiche de prononciation a tenté d'atteindre un degré de lisibilité maximale. Sur 20 informateurs, 13 ont effectivement lu la fiche. Parmi eux figurent 11 femmes et 2 hommes (âgés de 26 et 11 ans). Ils avaient été scolarisés en français. Tous ont très vite repéré et clairement signalé les faits d'écriture ou de prononciation méritant réflexion. C'est justement la participation active des informateurs, leurs suggestions face à une proposition graphique qui a permis d'évaluer la fiabilité et le degré de cohérence de la méthode d'investigation. Face à des réactions que la conception première de l'enquête n'avait pas voulu intégrer, il était donc légitime de séparer deux sortes de résultats. Les résultats srtictement prosodiques relevant de l'analyse de la durée, de l'intensité, de la hauteur et de la mélodie des traits articulatoires. Les observations liées à la perception et à la production de l'activité langagière appréhendée en relation avec les structures formelles et sémantiques de la langue considérée. Compte-tenu du support visuel choisi, il importait également de se demander en quoi la scripturalité du support conditionnait le décodage des mots proposés.

En effet, le locuteur avait sous les yeux 3 fiches de mots présentés en colonnes et placés les uns sous les autres. Le regard se déplaçait de gauche à droite sur l'axe horizontal et de haut en bas sur l'axe vertical. Le locuteur passait d'un mot court (monosyllabique) à un mot plus long (dy-, puis polysyllabique). Il allait aussi d'une structure simple à une structure plus complexe. Le troisième volet qui regroupait les polysyllabiques et les synthèmes, a été plus facile à lire (ex: larivyèr "rivière", lapouddmaniòkk "amidon de manioc", lapaykann' "feuille de canne à sucre", löronm' arranzé "rhum aromatisé", kari massalé "plat cuisiné", kari baton mouroungg "plat cuisiné").

Le choix du critère de simplification (matérielle et structurelle) explique le regroupement des items selon leur appartenance à une même classe sémantique (fruit, légume etc...). Cette technique classique a crée un effet de familiarisation permettant à l'informateur de se sentir en terrain connu, c'est-à-dire proche de sa réalité quotidienne. Tous ont très vite repéré et indiqué les notations qui leur posaient problème, car pouvant se prononcer de manière différente dans la communauté (ex: somiz, sëmiz, chömiz pour "chemise"). On remarque que la prononciation d'une voyelle ou d'une consonne peut intervenir au cours de la lecture à haute voix, car il arrive que l'informateur donne son sentiment sur l'écriture d'un son, sans avoir conscience que lui-même lit à haute voix non pas ce qui est écrit mais ce que lui dicte sa prononciation individuelle. Ex: *goni* "toile de jute" a été lu *gouni*. Le sentiment du locuteur est une variable socio-linguistique qui n'est pas fiable dans ce genre de description. L'étude phonologique en revanche avait bien indiqué la dynamique de l'opposition /o/~/u/ (Staudacher-Valliamee, G., 1991a). L'enquête montre par ailleurs que dans ce contexte, le décodage du mot écrit constitue un processus global en cela que les lecteurs identifient d'abord les signes écrits d'un signifié qui connote un référent culturel. Puis ils évaluent la forme du signifiant (écriture du mot). Deux types de structures conditionnent visiblement le mécanisme de la lecture: le caractère linéaire de la langue écrite ainsi que la très grande fréquence, dans le lexique réunionnais, de synthèmes ou de combinaisons d'unités dont le sens global n'est lisible qu'à la fin. Un informateur de 26 ans expliquait qu'il s'aidait de la fin du mot pour comprendre tout un synthème. Il semble aussi que l'origine du mot joue un rôle important. Les mots en j et en w illustrent bien ce phénomène. Dans des mots d'origine tamoule (rougay), indo-portugaise (pipangay) et malgache (makatja) le choix des lettres j et w n'a pas gêné la lecture alors que dans des mots d'origine française (sitrouij, bouj, grijé, anwi, kwi), des difficultés sont apparues. Ex: *kwi* se lit moins facilement que *kui* pour "cuit", *anwi* se lit moins bien que *annui* pour "agace". L'écriture de la consonne longue à l'aide d'une double consonne ne pose aucun problème en finale de mot (ex: *atann* "attendre", *sipèkk* "chipie"). A l'intérieur du mot, elle se lit difficilement quand il y a déja une voyelle nasale (ex: *ban'na* "ils", a été lu "banane"). Le déroulement des actes de lecture éclaire le fonctionnement de l'activité langagière chez les locuteurs créolophones tout en renseignant aussi sur certains mécanismes cognitifs en contexte plurilingue où les combinaisons de phonèmes ainsi que la structure des mots sont unifiés par les habitudes articulatoires. C'est cette oralité qui conditionne manifestement la lisibilité des items. Le traitement phonologique du lexique montre à quel niveau le français, langue officielle de l'écrit et langue première en milieu scolaire, peut interférer dans la lecture d'une partie du lexique chez certains lecteurs créolophones. dans l'enquête. Il apparaît que les mots d'origine non française peuvent couler une écriture créole (ex: pwakk, goni, kari, brèdd, soinj̃, gainj̃) sans freiner le décodage du français, car ils participent au découpage de la réalité insulaire par les locuteurs. L'inflexion sur la voyelle, l'accent circonflexe, l'accent grave, l'accent aigu sont utilisables en tant que signes graphiques pour compléter les lettres. Ces diacritiques ont une valeur pédagogique dans la mesure où ils visualisent les pertinences articulatoires du système créole. L'étude phonologique avait noté toutes les consonnes longues en finale absolue à l'aide des deux points après la consonne (ex: /patat:/ "patate douce, rhizome"). La fiche de prononciation avait adopté la régle d'écriture de la double consonne pour la sifflante sourde longue (ex: kalbass). Or, l'enquête prosodique révèle que toutes les consonnes (orales et nasales) ne peuvent être traitées de la même façon.

CONCLUSION

La lecture à haute voix, quand elle est ainsi testée à l'aide d'outils et de méthodes dictés par l'application des principes de phonologie précise les critères de lisibilité en contexte oral et écrit. On avance ainsi vers une étude de l'oralité créolophone dans son rapport au contexte. On a pu observer sur le terrain que les besoins des locuteurs créolophones dépendent de leur âge, de leur profession et du degré d'exposition à d'autres langues. Les enfants de 11 ans lisent très vite. L'apprentissage ne rencontre pas les mêmes difficultés suivant qu'il est destiné à des adultes ou à de jeunes enfants. Dans ce domaine, il serait imprudent de se contenter d'une fiche de mots soumis uniquement à la lecture à haute voix. Il faudrait étendre l'observation à la lecture (silencieuse et à haute voix) de véritables textes écrits. Cependant, la prise en compte de toutes les réalités du mot en tant qu'unité distinctive s'avère utile pour une première approche de l'activité langagière dans sa relation à l'architecture cognitive du terrain créolophone. L'enquête réalisée sur les bases d'une étude prosodique, confirme le primat de la tradition orale dans l'activité de lecture en créole. Ces résultats livrent un matériau de base pour une analyse cognitive de la lecture et de l'écriture dans le miroir des actes de paroles. Son apport complèterait les travaux réalisés et contrôlés en laboratoire sur la reconnaissance des mots (Morais, J., 1993).

BIBLIOGRAPHIE

Baddelay, A, 1993, *La mémoire humaine. Théorie et pratique,* P.U.G, 547 p.

Morais, J., 1993, Composantes lexicales et prélexicales de la reconnaisssance des mots in Actes du Colloque Interdisciplinaire du Comité national de La Recherche Scientifique *Images et langages. multimodalité et modélisation cognitive* 1-2 avril 1993 Paris. CNRS.

Staudacher-Valliamee, G., 1989 a, *Phonologie du créole réunionnais: unité et diversité,* Thèse de Doctorat Nouveau Régime soutenue en Juin 1989 à l'Université de Paris V.

—1989 b, Faits phonologiques et faits lexicaux dans le créole réunionnais, *Actes du 16e Colloque International de Linguistique Fonctionnelle,* Paris, Sorbonne, 29 Juin - 4 Juillet, pp 107-111.

—1990, Le système consonantique du créole réunionnais, *Cahiers du LACITO, Revue d'Ethnolinguistique,* N°5, pp. 89-108.

—1991 a, La dynamique du créole réunionnais: une synchronie au service de la diachronie, Actes du 20e Colloque de Linguistique Fonctionnelle, Prague, 11-17 Juillet, pp. 150-155.

—1991 b, Le vocalisme créole réunionnais, *Cahiers du LACITO, Revue d'Ethnolinguistique,* N°6, pp 20-37.

—1992 a, La phonologie du mot en créole réunionnais, Communication au XVe Congrès International des Linguistes, Québec, Canada, 9 - 14 Août 1992, *Actes du C.I.L. Tome 2.*

—1992 b, *Phonologie du créole réunionnais: unité et diversité,* Paris, Peeters/Selaf, 365, 190 p.

—1993, Die synchron dynamische Phonologie des Réunion Créole als Ausgangspunkt zur Annäherung an Kreolisierung und Sprachwandel, in Creolization and language changes, Düsseldorf: Niemeyer, à paraître.

Valdman Albert, 1977, Le créole: langue, statut et origine, Paris, Klincsieck.

LE ROLE DES SIMILARITES ACOUSTICO-PHONETIQUES DANS L'ACCES AU LEXIQUE CHEZ LE BILINGUE ARABE/ANGLAIS : PREUVES A PARTIR DU GATING

BOUDELAA, Sami
LABORATOIRE DE PHONÉTIQUE, U. F. R. L.
UNIVERSITE PARIS 7

Résumé

Le paradigme du "gating" est utilisé pour étudier l'effet des similarités acoustico-phonétiques sur l'accès au lexique chez le bilingue arabe-anglais (ie. le sujet qui peut utiliser plus d'une langue en cas de besoin). L'emploi d'une liste mixte de mots montre que le bilingue a besoin de plus d'input acoustique pour isoler et identifier un mot non-marqué (qui contient des segments similaires dans les deux langues) qu' un mot marqué (qui contient des segments spécifiques à l'une des deux langues). Le même effet est obtenu avec une liste bloquée par langue (ie. qui ne contient que les items d'une seule langue). Les données révèlent aussi une différence significative du même ordre entre le bilingue et les monolingues Arabes et Anglais. Les données sont discutées à la lumière des hypothèses de l'association lexicale et de la médiation conceptuelle.

Mots-clés: accès au lexique, bilinguisme, représentation lexicale, similarités acoustico-phonétiques.

INTRODUCTION

La présence de propriétés formelles communes à deux langues données influence les différentes étapes du traitement lexical chez le "bilingue", celui le sujet qui peut utiliser plus d'une langue en cas de besoin. Cependant, la manière dont cette influence agit sur le processus d'accès au lexique et sur la configuration cognitive du bilingue est controversée. Deux hypothèses se proposent d'expliquer la nature et l'effet des liens sous-jacents aux lexiques bilingues.

Selon la première hypothèse, celle de l'association lexicale les liens se font au niveau lexical. L'accès au sens d'un mot de la langue B se fait toujours par le biais de son synonyme dans la langue dominante A (Magiste, 1986., Gekoski, 1980). Certaines données empiriques supportent cette position. Kolers (1966) a montré, par le biais d'une tâche de rappel libre, que la présentation des équivalents de traduction comme "tree" suivi plus tard par "arbre" dans une liste mixte améliore le rappel autant que la simple répétition d'un même mot.

Selon la deuxième hypothèse, celle de la médiation conceptuelle les liens hypothétiques interlangues se font via une représentation conceptuelle indépendante de la variable langue. Le bilingue peut opérer dans l'une de ses deux langues sans qu'il y ait de contact avec les eprésentations lexicales de l'autre langue. Kirsner et al., (1984), dans une tâche de décision lexicale, montrent qu'il y a une facilitation interlangue si les sujets sont délibérément chargés de traduire les stimuli lors de leur première apparition. Un effet de facilitation inter-langue est aussi observé quand on présente aux sujets des équivalents de traduction homographes (Cristoffanini et al., 1986). En outre, Beauvillain et Grainger (1987); Grainger et Beauvillain, (1988) utilisant un paradigme d'amorçage, constatent que la séquence graphique "PAIN" facilite aussi bien le mot français "beurre" que le mot anglais "ache" pourvu que la présentation de la cible suive immédiatement l'amorce; le cas échéant seule la cible reliée à l'acception la plus fréquente du mot amorce est facilitée. Ces résultats laissent penser que l'accès au lexique chez le bilingue est non-sélectif pourvu que les stimuli partagent la même représentation graphique. Cette facilitation interlexicale est considérée comme relevant de l'aspect automatique des processus perceptifs.

Partant de l'idée selon laquelle une séquence graphique non-marquée comme "COIN" ou "PAIN" qui existe en français et en anglais peut donner lieu à un accès non-sélectif (Beauvillain et Grainger, 1987), on va essayer de savoir si une séquence de sons non-marqués peut susciter un tel accès. En d'autres termes, on va étudier le rôle que joue la similarité acoustico-phonétique au cours de la reconnaissance du mot chez le bilingue arabe/anglais. A ce propos, il faut noter que selon Altenberg et Cairns (1983) des variables infralexicales, comme les contraintes phonotactiques, sont déterminantes au cours du processus d'accès lexical chez des bilingues anglais/allemand. Ces auteurs ont présenté leur stimuli visuellement. Mais cette modalité ne favorise pas l'étude de l'effet des contraintes phonotactiques, mais pluôt l'effet des contraintes graphotactiques.

Grosjean (1984) a utilisé des logatomes pour étudier l'effet de la spécificité linguistique des sons sur l'accès lexical chez des bilingues français/anglais. Plus la cible possède des sons semblables dans les deux langues, plus la reconnaissance est lente. L'utilisation des logatomes rend difficile l'interprétation de leur résultats. On ne sait pas si c'est l'effet de la non-spécificité du signal qui retarde la décision des sujets ou c'est la non-lexicalité du matériel expérimental.

Afin d'éviter l'effet de tels facteurs on a utilisé la tâche du "gating" avec des mots réels pour voir si la variable infralexicale "spécificité du son" a une incidence sur le processus d'accès au lexique chez le bilingue arabe/anglais. Les deux hypothèses esquissées font les prédictions suivantes quant au rôle des similarités acoustico-phonétiques: l'hypothèse de l'association lexicale qui veut que l'accès au sens d'un mot de la langue B se fait toujours par le biais de son synonyme

de la langue A, impliquerait que le bilingue arabe/anglais aurait besoin de plus d'input acoustique pour la reconnaissance d'un mot anglais que d'un mot arabe. Cette hypothèse prédit aussi une différence entre les quantités d'input dont le bilingue et le monolingue anglais auront besoin pour reconnaître une cible quelconque, tandis qu'aucune différence n'est prévue entre les quantités d'input dont le bilingue arabe/anglais et le monolingue arabe auraient besoin.

L'hypothèse de la médiation conceptuelle qui veut que le bilingue puisse fonctionner dans l'une de ses deux langues sans aucun contact avec l'autre prédit que la variable mot-marqué serait opérationnelle seulement si on présente au bilingue des listes mixtes contenant les items des deux langues. En d'autres termes, il n' y aurait pas de différences entre les bilingues opérant avec des listes contenant les items d'une seule langue et les deux groupes de monolingues.

EXPÉRIENCE

METHODE

Matériel: Le corpus contient 60 mots qui sont répertoriés de la manière suivante:

-30 mots dissyllabiques non-marqués dont l'agrégat consonantique interne est commun aux deux langues. e.g. anglais:"basket", arabe: /næskot/ (nous nous taisons).

-15 mots dissyllabiques dont l'agrégat consonantique interne est spécifique à l'anglais. e.g: "depress".

-15 mots dissyllabiques dont l'agrégat consonantique interne est spécifique à l'arabe. e.g: /mæl?æb/ (un stade).

Les mots non-marqués respectent les règles phonotactiques des deux langues, contiennent des sons communs aux deux langues (Flege et Port, 1981) et peuvent avoir la même frontière syllabique. Les marqués contiennent des segments qui appartiennent exclusivement à une seule langue (eg. sholplift, /mæl?æb/ stade). Tous les mots ont la même fréquence d'usage.

Locuteur: Un locuteur bilingue a été choisi pour enregistrer le corpus. Selon deux sujtes natifs de l'arabe te l'anglais son accent n'est pas marqué dans aucune des deux langues.

Procédure d'enregistrement: L'enregistrement est fait dans une chambre sourde en utilisant une cassette (UX 60) et un magnétophone (Radiola D 69 20MT2). Le locuteur était placé devant le microphone à une distance d'environ 25 cm. Il devait lire des listes mixtes sur lesquelles chaque mot arabe est suivi d'un mot anglais. Il s'est familiarisé avec les listes avant de commencer l'enregistrement. Les mots ont été présentés isolément afin d'éviter les effets de coarticulation et d'abrègement de la durée du signal. Le corpus a été ensuite enregistré sur disquette par le biais d'un Macrecorder pour y effectuer le procédure du "gating" à l'aide d'un logiciel Soundedit T.M. (version 1989 2.0.1.). 50 ms du signal sont d'abord sélectionnés puis enregistrées via un cable "RCA Mini-Jack" sur une cassette [UX 60] en utilisant un magnétophone NEC RT 320. L'intervalle entre les "gates" est de 10 secondes, tandis que celui entre les mots est d'environ 50 secondes. Des distracteurs ont été introduits aléatoirement de façon à obtenir en moyenne un distracteur tous les trois mots cibles.

Sujets: La passation du test a eu lieu individuellement et a duré environ 30 minutes. 20 sujets y ont participé:

a/10 sujets bilingues arabe-anglais dont l'âge moyen est de 27 ans. Ils ont tous une expérience de dix ans au moins de l'anglais.

b/ 5 sujets monolingues anglais dont l'âge moyen est de 28 ans.

c/ 5 sujets monolingues arabes dont l'âge moyen est de 27 ans.

Plus des 85% de ces sujets ont déclaré avoir une expérience antérieure du français, mais qu'ils ne se considèrent pas comme des bilingues arabe-français dans des arabophone, ou des bilingues anglais-français dans le cas des monolingues anglais. Par ailleurs, ils ont tous une ouïe normale.

Procédure: Avant la présentation auditive du matériel expérimental des écouteurs (EM 6126), les sujets se sont familiarisés avec la tâche à l'aide d'un bloc d'entraînement; les stimuli étaient présentés à un niveau d'intensité jugé confortable par les sujets.

5 sujets bilingues ont écouté deux listes de stimuli, A et B, dont chacune contenait 30 mots (listes bloquées par langue). La liste A contenait 30 mots arabes dont 15 sont marqués et 15 non-marqués. La liste B contenait 30 mots anglais dont 15 marqués et 15 non-marqués.

L'ordre de présentation des stimuli arabe/anglais, marqués/non-marqués a été déterminé de façon aléatoire pour les 5 autres bilingues qui ont écouté les 60 mots (listes mixtes). Les monolingues arabes et anglais ont entendu les 30 mots de leur langue respective.

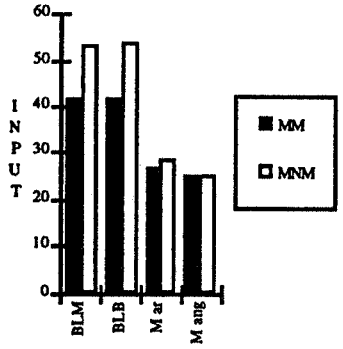
Les mots sont présentés en segments de longueur croissante. Le sujet n'entend d'abord que les 50 premières ms du mot et doit noter par écrit le mot qui correspond le mieux à l'input qu'il a perçu. Ensuite, les 100 premières msc sont présentées et le sujet donne une nouvelle réponse et ainsi de suite jusqu'à ce que le mot en entier ait été présenté. Le sujet doit également exprimer son degré de certitude quant au mot qu'il propose en mettant une croix dans l'une des quatre cases d'une échelle allant de "pas sûr du tout" à "très sûr". Par ailleurs, le protocole expérimental de la liste bloquée par langue est associé à un mode monolingue et celui de la liste mixte à un mode bilingue (ces termes sont utilisés de façons équivalentes).

RÉSULTATS

Trois types d'analyse sont faits: le premier comprend l'analyse du point d'isolement (PI) et du point de reconnaissance (PR) du mot chez les bilingues dans les deux modes et chez les monolingues. Il consiste à décrire, comparer et expliquer les différences entre ces groupes de sujets. Le second vise à sonder le degré de certitude des sujets concernant leurs suggestions au PI. Le dernier porte sur le processus de reconnaissance du mot.

LE P.I. DU MOT

Le terme PI réfère au point où le sujet propose le mot correct sans changer d'avis ultérieuement. Le graphique (1) montre la quantité moyenne d'input dont les quatres groupes de sujets ont besoin pour isoler un mot quelconque. Afin de calculer le PI de chaque item la fenêtre moyenne à laquelle les sujets ont proposé le mot correcte sans changer d'avis ultérieurement a été localisée.



Graphique (1): Quantité d'input acoustique (en %) nécessaire pour isoler un mot marqué (MM) et un mot non-marqué (MNM) chez le bilingue opérant en liste bloquée (BLB), le bilingue opérant en liste mixte (BLM), chez le monolingue arabe (M ar) et le monolingue anglais (M ang).

On a soumis les valeurs des PI à une ANOVA à deux facteurs avec (4 groupes de sujets x 2 types d'items). Globalement il y a un effet du facteur groupe de sujets [$F(3,1) = 59,37$, $p<0,0001$], en ce sens que les sujets manifestent des comportement différents selon qu'ils sont bilingues ou monolingues. Il y a aussi un effet du facteur item [$F(3,1)= 24,52$, $p<0,0001$], ce qui implique que la nature du mot à identifier joue un rôle. L'interaction entre ces deux facteurs est significative [$F(3,1)= 3,92$ $p<0,01$]. Toutefois, une analyse statistique plus détaillée révèle que les bilingues en mode monolingue et ceux en mode bilingue traitent les stimuli arabes et anglais de la même façon, et que leurs sensibilités aux MM/MNM sont semblanbles. [test de Scheffe, $p = o,5$]. Les deux groupes de bilingues sont différents des monolingues anglais et arabes [Scheffe F-test, $p<0,05$].

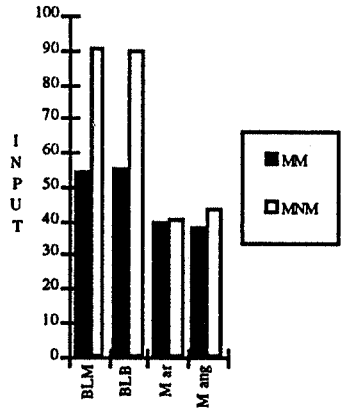
Discussion

Les résultats de cette expérience suggèrent que la quantité d'input acoustique nécessaire pour isoler un mot quelconque est sensible à la spécificité des sons qui le composent et au type d'auditeur. Les bilingues sont aussi différents des monolingues arabes que des monolingues anglais ce qui contredit les prédictions de l'hypothèse de l'association lexicale qui prédit l'existence d'une différence entre les performances des bilingues et des monolingues anglais ainsi que l'absence d'une telle différence entre les bilingues et les monolingues arabes. L'absence d'effet de la variable MM/MNM chez les monolingues est essentiellement due au fait que chez ces sujets le processus d'accès au lexique concerne un seul dictionnaire mental. Il en va tout autrement pour le bilingue qui s'avère sensible à cette variable puisque ses connaissances linguistiques mettent à sa disposition tous les mots conformes à l'input acoustique perçu. Ceci reflète l'existence d'un accès initialement neutre dans la mesure où le stimulus ne permet pas au sujet de décider de la langue d'appartenance de l'item à identifier.

Par ailleurs, les bilingues opérant en mode monolingue devaient, selon l'hypothèse de la médiation conceptuelle, avoir besoin de moins d'information acoustique pour reconnaître les mots que les bilingues opérant en mode bilingue puisque les premiers étaient quasiment sûrs de la langue d'appartenance de l'item à reconnaître. Tel n'est pas le cas ce, qui pourrait signifier que le fait que le bilingue sait qu'il a affaire à une liste bloquée de sa langue A n'empêcherait pas qu' un mot de sa langue B soit activé à condition que ce mot soit compatible avec l'input perçu. C'est dans ce sens là qu'il faut comprendre l'idée selon laquelle le bilingue n'est jamais l'ensemble de deux monolingues puisque même en fonctionnant dans sa langue A, sa langue B est toujours relativement activée.

LE P.R. DU MOT

Le point de reconnaissance est le point où les sujets étaient sûrs à 100% du mot qu'ils proposaient. Le graphique (2) montre la quantité moyenne d'input acoustique nécessaire pour la reconnaissance d'une cible donnée chez les quatre groupes de sujets.



Graphique (2): La quantité d'input acoustique (en %) nécessaire pour reconnaître les MM et MNM chez le BLB, le BLM, le M ar et le M ang.

Les résultats de l'ANOVA montrent que l'effet du facteur groupe de sujets est significatif [$F(3,1) = 66,47$, $p< 0,0001$]. Le facteur items est opérationnel aussi [$F(3,1) = 71,51$ $p<0,0001$]. Ces deux facteurs interagissent d'une manière significative [SxI, $F(3,1)=18,93 p< 0,0001$], impliquant ainsi que la performance des sujets varie en fonction des deux facteurs principaux. L'analyse post hoc révèle qu'il n' y a pas des différences significative entre les deux groupes des bilingues [Scheffe F-test, $p=0,5$] et qu'ils sont, tout les deux sensibles seulement à la variable MM/MNM [test de Scheffe, $p<0,05$]. On observe une différence

significative entre les deux groupes de bilingues et les deux groupes de monolingues [test de Scheffe, p<0,05].

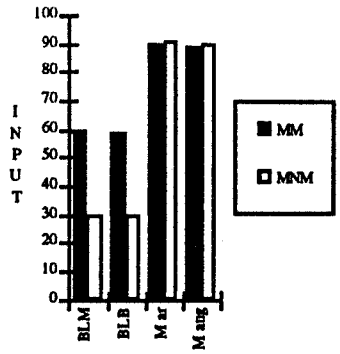
DISCUSSION

Ce résultat permet de dire que la reconnaissance de mots isolés par des sujets bilingues est sensible à la spécificité des sons composant le mot et non à la variable langue en tant que telle. Si le sujet est bilingue, et si la cible qu'on lui présente ne contient pas d'indices acoustico-phonétiques spécifiques, celui-ci aura besoin de plus d'informations pour la reconnaître. Les bilingues se comportent de la même manière en mode bilingue et en mode monolingue en ce sens qu'ils ont besoin de plus d'input pour recnnaître un mot quelconque que les deux groupes de monolingues. Ce résultat contredit l'hypothèse de l'association lexicale puisque la langue ne semble pas être une variable pertinente et que les bilingues se comportent différemment des monolingues anglais et arabes. Un bilingue à qui on présente un input conforme aux règles de ses deux langues se trouve face à un nombre plus important de candidats qui sont activés en harmonie avec l'input acoustico-phonétique perçu. Cette opération d'activation plus exhaustive retarde, vraisemblablement la reconnaissance de la cible non-marquée.

Plus important encore, la variable liste mixte/liste bloquée n'a pas d'incidence sur les performances des bilingues; cela infirme l'hypothèse de la médiation conceptuelle. Ceci implique que le coût de traitement de l'information chez le bilingue est fonction de la spécificité linguistique qui caractérise le mot à reconnaître puisque, même quand le bilingue est en mode monolingue, des items appartenant aux deux langues s'activent chez lui. Tout se passe comme si, au cours du traitement d'un mot isolé, une unité infralexicale non-marquée comme la fricative /s/ activait tous les mots qui commencent par /s/ indépendamment du fait que ce mot appartienne au lexique arabe ou au lexique anglais, d'où ce besoin d'un input plus important chez les bilingues dans les deux modes linguistiques.

LE DEGRÉ DE CERTITUDE DES SUJETS AU "PI" DU MOT.

Pour étudier le degré de certitude chez les sujets, on a procédé de la façon suivante: Si le sujet propose un mot "X" et met une croix dans la première case, il est supposé être sûr à 12,5% de sa proposition; si la croix est mise dans la deuxième case, son degré de certitude est de 25%. Ainsi, on a pu assigner un degré de certitude à chaque sujet pour chaque mot. Le graphique (3) montre le degré de certitude des quatre groupes de sujets.



Graphique (3): le degré de certitude au PI des MM et des MNM chez le BLB, le BLM, le M ar et le M ang.

Les données ont été soumises à une ANOVA à deux facteurs. Le facteur groupe de sujets joue un rôle en ce sens que les sujets diffèrent globalement quant à leur degré de certitude [F(3,1) = 8,43, p<0,0003]. L'effet du facteur item est opérationnel [F(3,1) = 16,5, p< 0,0003]. En outre, les deux facteurs interagissent [SxI, F(3,1) = 5,69, p<0,0003). Cependant, l'analyse post hoc révèle que les deux groupes des bilingues se comportent de la même façon vis à vis des MM/MNM [Scheffe F-test, p= 0,5]. Il y a une différence significative entre les deux groupes de bilingues et les deux groupes de monolingues [Scheffe F-test, p<0,05]; ces derniers étant plus sûrs.

DISCUSSION

Si l'on admet que, chez le monolingue la reconnaissance du mot est un processus pendant lequel l'information sensorielle détermine un ensemble de candidats qui entrent en compétition pour que l'un d'entre eux soit reconnu, on peut dire que, schématiquement, deux opérations régissent le processus de reconnaissance du mot chez le monolingue: une opération d'élaboration et une autre de sélection; mais chez le bilingue une autre opération de sélection intervient: le choix entre l'une des deux langues. Ainsi, le bilingue commence le processus de reconnaissance du mot avec une tâche supplémentaire par rapport au monolingue au sens où il doit opérer un choix concernant le lexique auquel il doit accéder (Grainger, 1993). Face à un item non-marqué, le bilingue ne peut pas entreprendre ce choix élémentaire et par conséquent ne peut pas accéder à une langue au détriment de l'autre, d'où ce faible degré de certitude qu'il manifeste par rapport au monolingue. Mais, si ceci est vrai, alors pourquoi les bilingues en mode monolingue ont-ils manifesté des degrés de certitude faibles comparable à ceux en mode bilingue?

On peut répondre à cette question en se référant à la Cohorte (Marslen-Wilson et al., 1987, 1991). Ces auteurs supposent qu'il y a trois niveaux d'organisation (le niveau trait, le niveau phonème et le niveau mot) et que le processus d'accès au lexique est autonome au cours des 300 premières ms. Appliqué au sujet bilingue le modèle de la Cohorte dirait donc que le bilingue dispose des deux lexiques interconnectés à différents

niveaux de représentation. Ainsi, outre les liens inter-niveaux intralangues (ex: les liens entre le niveau trait et le niveau phonème de la langue arabe), il y aurait des liens *interniveaux interlangues* dans le sens où il y aurait des liens entre le niveau trait du lexique arabe et le niveau trait du lexique anglais, entre le niveau phonème en arabe et sa contrepartie anglais, etc... La force des liens interniveaux interlangues est déterminée par le degré de similarité entre les deux langues du bilingue.

Si l'on suppose admis le cadre théorique de la Cohorte appliqué au bilingue, on pourrait imputer la faiblesse des degrés de certitude chez les deux groupes de bilingues aux liens interniveaux interlangues qui sous-tendent un processus *d'irradiation de l'activation irrépressible* entre les deux lexiques du bilingue. Le fait qu'un mot quelconque contient des unités infralexicales communes aux deux langues aurait pour conséquence de déclencher une activation initialement neutre et quasi-équivalente des deux lexiques du bilingue. Ce faisant, les bilingues en mode bilingue et ceux en mode monolingue ne peuvent pas être assez sûrs de la langue d'appartenance de l'item présenté puisque la structure infralexicale engendre un niveau d'activation équivalent dans leurs deux lexiques.

LE PROCESSUS DE LA RECONNAISSANCE DU MOT

On suppose ici, à l'instar de Grosjean (1980), que les mots proposés par les sujets sont similaires aux items qui seraient activés au cours du traitement "en ligne".

Nos données infirment les prédictions de l'hypothèse de la médiation conceptuelle puisque les deux groupes de bilingues ont proposé des mots arabes et des mots anglais au cours de la reconnaissance de la cible "basket". Les bilingues en mode monolingue ont suggéré en moyenne 48 à 51% des items qui appartiennent à la langue non-pertinente. Ceux en mode bilingue ont proposé des cohortes contenant 49 à 52% des mots de la langue non-pertinente. Afin de s'assurer que les compositions des cohortes proposées par les deux groupes de bilingues ne diffèrent pas d'une manière significative, les données ont été soumises à une série d'analyses statistiques en procédant de la manière suivante: à chaque mot X proposé au cours de la reconnaissance d'une cible donnée on a assigné la valeur 2 si le mot appartient à la même langue que la cible en cours de présentation et la valeur 1 dans le cas contraire. Les valeurs ainsi obtenues ont été soumises à un test statistique non-paramétrique Mann-Whitney (U) qui a révélé qu'il n' y a pas de différence significative entre les deux groupes de bilingues ($U(30) = 57,5$, $p = 0,6$). Ce résultat suggère qu'au cours du processus de la reconnaissance du mot les deux lexiques des bilingues, en mode monolingue et en mode bilingue, sont activés d'une manière quasi-équivalente dans la mesure où les mots des deux langues ont des chances équiprobables d'être activés. Les unités infralexicales peuvent à elle seules déclencher un processus d'accès non-sélectif, ce qui implique qu'au cours de la reconnaissance des mots isolés par des bilingues les liens interlexicaux hypothétiques prennent essor et deviennent aussi puissants que les liens intralexicaux; d'où cette cohorte mixte des mots non-marqués.

Les cohortes proposées par les bilingues en mode bilingue pour les mots marqués contiennent en moyenne 80% des items appartenant à la langue pertinente. Par exemple, pour la cible arabe marquée /muqliq/ (ie. ennuyeux) les bilingues en mode bilingue ont suggéré deux mots anglais sur une cohorte de huit items. Les cohortes de bilingues en mode monolingue contiennent une faible nombre d'items de la langue non-pertinente, soit 21%. Le test Mann-Whitney révèle l'absence de différence significative entre les cohortes proposées par les deux groupes de sujets pour les items marqués ($U(13) = 35$, $p = 0,9$). Ce resultat reflète l'absence de l'effet de la variable liste bloquée/liste mixte et infirme par conséquent l'hypothèse de la médiation conceptuelle.

Les cohortes monolingues sont différentes de celles des bilingues. Les monolingues arabes et anglais ont proposé des mots que les bilingues n'ont pas proposés infirmant ainsi l'hypothèse de l'association lexicale puisqu'elle prédisait une différence significative entre les bilingues et les monolingues anglais. Tout semble se passer comme si les deux langues n'étaient jamais complètement désactivées chez le bilingue (Grosjean, 1992).

CONCLUSION GÉNÉRALE

Nos résultats contredisent non seulement l'hypothèse de l'association lexicale mais elles infirment aussi un des fondement de l'hypothèse de la médiation conceptuelle concernant l'accès au lexique chez le bilingue: selon celle-ci le bilingue pourrait opérer dans un mode totalement monolingue. Il s'est avéré que l'accès au lexique non-sélectif peut être observé non seulement chez des bilingues en mode bilingue mais aussi chez le bilingue en mode monolingue pourvu qu'il y ait des unités similaires dans ses deux langues. Ce type d'accès est déclenché d'une manière plus ou moins évidente selon que la cible à reconnaître est marquée ou pas. Il semble par ailleurs, que les deux niveaux de représentation postulés par les hypothèses de l'association lexicale et de la médiation conceptuelle sont insuffisants. Il est plus fructueux de postuler l'existence de plusieurs niveaux de représentation comme le trait, le phonème et le mot et de dire que des liens interlexicaux interniveaux s'instaurent en fonction des similarités sous-jacentes aux deux langues.

BIBLIOGRAPHIE

Altenberg, E & Cairns, H. (1983). The effects of phonotactic constraints on lexical processing in bilinguals and monolingual subjects. *Journal of Verbal Learning and Verbal Behavior.* 22, 174-188.

Beauvillain, C. & Grainger, J. (1987). Accessing interlexical homographs: Some limitations of language-selective access. *Journal of Memory and language.* 26, 658-672.

Cristoffanini, P., Kirsner, K. & Milech, D. (1986). Bilingual lexical representation: The status of Spanish-English cognates. *The Quarterly Journal of Experimental Psychology*. 38A, 367-393.

Flege, J & Port, R. (1981). Cross phonetic interference: Arabic to English. *Language and Speech.* 24, 125-146.

Gekoski, W. (1980). Language acquisition context and language organization in bilinguals. *Journal of Psycholinguistic Research.* 9, 429-449.

Grainger, J. & Beauvillain, C. (1988). Associative priming in bilinguals: Some limits of interlingual facilitation. *Canadian Journal of Psychology* 42, 261-273.

Grainger, J. (1993). Visual word recognition in bilinguals. (in) R. Schreuder et B. Weltens (eds) *The Bilingual Lexicon.* Benjamin Publishing, Amsterdam.

Grosjean, F. (1984). Vers une psycholinguistique expérimentale du parler bilingue. (dans) G. Lüdi, *Devenir Bilingue, Parler Bilingue.* Verlag Tübingen.

Grosjean, F. (1992). Another view of bilingualism. (in) R, Harris, *Cognitive Processing in Bilinguals*, Elsevier.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics.* 28, 267-283.

Kirsner, K., Smith, M., Lockhart, R., King,. M. & Jain. M. (1984). The bilingual lexicon: Language specific units in an integrated network. *Journal of Verbal Learning and Verbal Behavior.* 23, 519-539.

Kolers, P. (1966). Interlingual facilitation of short-term memory, *Journal of Verbal Learning and Verbal Behavior.* 5, 314-319.

Lahiri, A. & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition,* 38, 245-294.

Magiste, E. (1985). Development of intra and interlingual interference in bilinguals: *Journal of Psycholinguistic Research*, 14, 137-154.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition.* 25, 71-102.

Remerciement:

Je tiens à remercier Mme Bacri, N, Mme Vaissiere, J., Mr Boulakia, G., Deschamps, A., Dommergues, J-Y. pour leur conseils. Je remercie aussi Mr Meftah, M, pour son aide.

VERS LA PERTE DE L'OPPOSITION DU LIEU D'ARTICULATION DES VOYELLES MOYENNES ARRONDIES DANS LA PAROLE DES JEUNES GENS D'ILE-DE-FRANCE

ISABELLE MALDEREZ

UFR LINGUISTIQUE, UNIVERSITE DENIS DIDEROT, PARIS 7.
10, RUE CHARLES V, 75004 PARIS; E-mail: isabelle.malderez@linguist.jussieu.fr

Résumé

Je présente ici les résultats d'enquêtes récentes réalisées dans le cadre d'une recherche sur la description acoustique des voyelles du français contemporain. Je ferai part plus précisément de postériorisation de réalisations de /ə/*.

J'ai réalisé une première enquête auprès d'enfants car les fautes d'orthographe d'enfants de 6-7 ans reflètent souvent leur conception phonologique. Ainsi, le relevé systématique de fautes lexicales atypiques devrait permettre de mettre en évidence les changements phonétiques récents. Les résultats de ce test confirment l'existence d'une tendance à la confusion des deux voyelles /ə/ et /O/ chez les jeunes enfants et permettent de dégager un contexte favorisant ce phénomène. Un second test, réalisé avec des adultes et concernant des mots à l'orthographe peu fixée, confirme les résultats précédents.

La seconde enquête concerne la réalisation de ces mêmes voyelles chez des adultes. Un test de perception détermine les réalisations déviantes de la voyelle /ə/ attendue. Ces réalisations sont ensuite décrites en termes de formants. Enfin, un traitement statistique met en évidence les facteurs linguistiques et extra-linguistiques favorisant la production postériorisée des /ə/.

Mots clés: *Changement phonétique, Socio-phonétique, Orthographe.*

INTRODUCTION

Cette communication présente les résultats de deux enquêtes récentes réalisées dans le cadre d'une recherche sur la description acoustique des voyelles du français contemporain (spontané et lecture) et en particulier du traitement des voyelles arrondies[1] en Ile-de-France. Je ferai part ici d'un phénomène peu mentionné à ma connaissance (Carton & al. 1983; 51), à savoir la postériorisation des réalisations de /ə/ dans un certain type de syllabes.

1. TESTS ORTHOGRAPHIQUES

1.1. PROLÉGOMÈNES

Il n'est pas question ici d'alimenter le débat sur la réforme de l'orthographe mais de présenter une procédure pour l'étude des changements phonétiques en cours. Il suffit de consulter la table des matières d'ouvrages scolaires spécialisés pour établir la listes des erreurs attendues: doublement ou non des consonnes, graphie des voyelles nasales, confusion des accents, omission ou présence d'un e, etc. Ces fautes sont liées à des changements phonétiques, en cours de réalisation ou achevés, relevés et étudiés depuis longtemps. Cependant, les enseignants et en particulier les instituteurs, sont souvent confrontés à des fautes d'orthographe lexicale atypiques - ignorées dans les exercices de régulation[2]. C'est le cas, par exemple, des troncations graphiques issues de la chute de /r/ ou de /l/ dans les groupes consonantiques finals (quate cartabes *quatre cartables*)[3].

Mon travail s'articule autour d'un double principe:

1- les enfants de l'école primaire, et en particulier les élèves des classes de CE1 et CE2, (âgés de 7-9 ans), écrivent souvent les mots tels qu'ils les prononcent;

2- à partir du CM1, (9 ans), on peut considérer que les enfants écrivent ce qu'ils ont appris et non plus simplement ce qu'ils "disent" - étant entendu que la limite entre les deux types de comportement n'est pas rigide et peut être avancée pour certains enfants et reculée pour d'autres.

Je considère, avec Fónagy (1971, 1989), que "les fautes d'orthographe d'enfants de 6-7 ans reflètent souvent leur 'conception phonologique'". Ainsi, le relevé systématique des fautes lexicales atypiques permet de mettre en évidence les changements phonétiques récents ou "en dessous du niveau de conscience sociale" (Labov 1976) . Si l'absence de fautes d'orthographe atypiques chez les enfants de plus de 9 ans ou chez les adultes ne nie pas la réalité d'un changement phonétique en cours, leur présence confirme son étendue.

1.2. DÉGAGEMENT D'UN ATYPISME

Lors d'un contact quotidien avec des enfants de 6-8 ans en qualité d'institutrice, j'ai relevé des réalisations[4] de /ə/ qui tendent vers [O] dans la parole d'un bon nombre d'entre eux. Ce déplacement se retrouve dans certaines réalisations graphiques. S.D., par exemple, fut très surpris quand je lui ai signalé l'orthoépie du mot qu'il avait écrit ropos. D'autre part, dans une classe de CM1 lors d'un contrôle de conjugaison, des enfants ont orthographié ils seront et ils eront les verbes *savoir* et *avoir* au futur.

Si le déplacement de /O/ vers [Ø] m'était déjà connu ([œ] chez Martinet 1969), le déplacement inverse me surprit. C'est pourquoi, afin de déterminer le degré de diffusion de cet éventuel changement phonétique en cours, je mène actuellement une étude systématique auprès d'enfants et d'adultes. Je présente ici les premiers

résultats de l'enquête orthographique menée dans trois écoles primaires rurales du sud de l'Oise[5].

1.3. PROCÉDURE

L'analyse de mon relevé de fautes atypiques m'a permis de dégager un contexte privilégié de cette variation:

/r/ — C_1^2 V (C) (V).

J'ai donc construit un premier test A composé de deux exercices de type scolaire et présentant principalement les variables /ə/ et /O/ dans le contexte favorisant le déplacement de /ə/ vers [O][6] et un second test B présentant le même type de contexte pour la paire /y/, /u/[7].

Le test A a été effectué par 230 enfants, de CE1, CE2, CM1 ou CM2, dans le cadre de leur classe durant les mois de décembre 1991 et janvier 1992, le test (b) par les enfants de CE1 uniquement en mars 1992. Les procédures de passation ont été identiques dans les trois écoles. Les consignes écrites furent expliquées oralement aux CE1 et CE2, les autres enfants étant supposés capables de se débrouiller seuls avec le texte. L'attention des enfants a été attirée sur le fait qu'il leur était permis de laisser un item sans réponse plutôt que de la donner au hasard.

1.4. RÉSULTATS GLOBAUX[8]

Si l'on considère l'ensemble des 230 enfants, le pourcentage global d'erreur au test est de 8,01%. Comme je m'y attendais, ce pourcentage décroît régulièrement de 16,39% au CE1 à 1,07% au CM2 (tableau 1). Il apparaît une diminution de 5% d'erreur d'un niveau au suivant. Les résultats de chaque paire de deux cours consécutifs sont statistiquement différents (One Factor Anova-Repeated mesures; Fisher PLSD=4,3; p≤ 0,02).

D'autre part, la répartition des erreurs sur les différents cours se décompose comme suit: CE1: 50%; CE2: 30%; CM1: 16%; CM2: 4%. Ces résultats restent à peu près constants (± 4% maximum) quand on considère les différents contextes ou les différentes voyelles. Ainsi 80% des erreurs sont attribuables aux enfants de 7-9 ans.

Les résultats sont sensiblement identiques en ce qui concerne le comportement des deux voyelles. Globalement le pourcentage d'erreur est le même pour les voyelles /O/ et /ə/ : 0,93 point de différence au CE1; 1,37 au CE2; 0,79 au CM1 et aucun au CM2 (tableau 1). Statistiquement, ce facteur n'est pas pertinent pour chacun des quatre niveaux (t-test non-apparié; d.l.=19; valeurs ce1=0,31 ce2=0,34 cm1=0,25 cm2=0; p(2-tail) ce1=0,75 ce2=0,74 cm1=0,8).

Le test montre d'autre part la pertinence du contexte:

/r/ — C_1^2 V (C) (V).

Pour chaque niveau, on trouve deux fois plus d'erreurs dans ce contexte que dans les autres contextes dans le cas des erreurs sur /ə/ - CE1: 1,96; CE2: 2,26; CM1: 1,98; CM2: 1,74 - et environ quatre fois plus d'erreurs dans le cas des erreurs sur /O/ - CE1: 4,01; CE2: 4,39; CM1: 4,02 (tableau 1).

Cependant, statistiquement, si le contexte ("r—" versus "$C_{[-r]}$—") est un facteur pertinent au CE1 et au CE2 (t-test non-apparié; d.l.=19; valeurs ce1=3,37 ce2=2,79; p(2-tail) ce1<0,004 ce2<0,02), il ne l'est plus au CM1 et au CM2 (valeurs cm1=1,58 cm2=1,76; p(2-tail) cm1>0,13 cm2>0,08).

Pour le test B, le pourcentage moyen d'erreur au CE1 n'est que de 11,67% bien que celui-ci présente des mots peu usités par les enfants. Statistiquement, les pourcentages d'erreur du test A sont significativement supérieurs à ceux du test B (test-t apparié, d.l.=49; Mean X-Y=6,48 [test1=18,32; test2=11,84]; t=3,96; p(1-tail) < 0,0001)[9].

1.5. LE CAS DE *MOTTE / MEUTE*

Cet item présente des résultats particuliers:

a) concentration des erreurs sur cet item dans les grandes classes (CE1: 9%; CE2: 16%; CM1: 33%; CM2: 69%) et pourcentage absolu d'erreur progressant du CE1 au CM1 (tableau 2); b) le plus grand pourcentage global d'erreurs par item: 47,51%; c) deuxième rang des réponses non exprimées au CE1; d) absence du contexte favorisant les erreurs (/r/ —).

J'ai donc supposé que *motte* ou *meute* était inconnu de la plupart des enfants testés. Les instituteurs, surpris par cette accumulation d'erreurs dans les grandes classes, ont vérifié la validité de cette hypothèse. Dans la plupart des cas la distinction sémantique de ces deux mots est neutralisée par assimilation: les enfants connaissent un élément *meu(ot)te* qui signifie "une certaine quantité de..., un regroupement de...".

1.6. DISCUSSION

Il apparait dans cette enquête qu'à l'écrit, en syllabe initiale, les voyelles e et o sont confondues par les jeunes enfants (7-9 ans), surtout dans le contexte dégagé ci-dessus:

/r/ — C_1^2 V (C) (V).

Cette confusion est plus importante que celle qui concerne la paire ou, u. Chez les enfants plus âgés, le pourcentage d'erreur est minime et le contexte n'est pas un critère opératoire (test A).

D'autre part, j'ai aussi pu mettre en évidence ce type de phénomène dans l'écrit d'adultes. J'ai effectué auprès de douze francophones natifs - étudiants de second et troisième cycle universitaire, âgés de 21 à 45 ans -, un court test C. Il s'agissait de retrouver la lettre manquante - la deuxième - de chaque mot d'une liste de dix[10]. Mon attention s'est portée sur le lexème *reblochon* qui est souvent prononcé [robloʃɔ̃]. On peut supposer que comme il s'agit d'un nom provenant d'une marque son orthographe n'est pas intégrée par tous. Sur l'ensemble des deux listes présentées à deux groupes distincts, j'ai obtenu 67% d'erreur: 83% pour la liste A, 50% pour la liste B (tableau 3). Dans cette population adulte le critère de l'âge n'est pas opératoire.

Ces premiers résultats ne sont pas encore tout à fait suffisants pour valider mon hypothèse initiale - mise en

évidence graphique des changements phonétiques. Les tests ne concernent que la paire (/ə/;/O/), et il n'est pas possible de comparer pour tous les niveaux ces premiers résultats avec ceux qui concernent la paire /u;y/, d'autres tests étant en cours de réalisation.

2- ETUDE DE REALISATIONS PHONETIQUES

La seconde expérimentation est constituée par une étude perceptive et acoustique de la réalisation de /ə/ dans la lecture oralisée des adultes. Le corpus[11] regroupe 6 phrases lues par 9 femmes et 7 hommes âgés de 20 à 55 ans. Je me suis attachée à étudier la voyelle se trouvant dans le contexte favorisant dégagé plus haut: *re* dans *reviendra(it)* et *recopiez*, ainsi que le /O/ de *recopiez* comme référence. Ici encore on trouve des cas de postériorisation dans les réalisations phonétiques de /ə/.

2.1. PROCÉDURE DE PASSATION

Le test de perception, réalisé par sept sujets entraînés - phonéticiens ou linguistes habitués à ce genre de tâche - a servi à déterminer les réalisations de la voyelle /ə/ attendue. Dans chaque phrase du corpus, j'ai isolé la syllabe cible, qui a été écoutée deux fois par le jury, sans qu'il ne sache de quel mot provenait le segment isolé. Il était alors demandé au sujet d'opérer un choix entre quatre réalisations possibles [ø œ o ɔ][12]- ou éventuellement d'en proposer une autre. Le jury a entendu les 111 syllabes[13] groupées par locuteur (93 issues de *reviendra(it)* et 18 issues de *recopiez*).

2.2. LA PERCEPTION DES /O/ ATTENDUS

Comme le montre le tableau 4, la majorité des /O/ attendus ont été perçus [ø] par le jury. Ceci va dans le sens de Martinet (1969). On peut noter ici que le phénomène touche aussi bien les hommes que les femmes, les plus jeunes que les plus âgés.

2.3. LA PERCEPTION DES /ə/ ATTENDUS

Quelquesoit le locuteur, les /ə/ du mot *reviendra(it)* sont toujours bien perçus [Ø][14] et ce avec des scores situés entre 71 et 100%. Le phénomène de postériorisation qui nous intéresse ici ne touche que le /ə/ du mot *recopiez*. C'est donc uniquement sur les réalisations de cet item que portent les analyses suivantes.

Globalement , c'est à dire si on considère les 16 locuteurs (18 occurences), la première syllabe de ce mot est perçue majoritairement[15] [Ø] dans 8 cas, majoritairement [O] dans 5 cas. Pour les cinq occurences réalisations, on obtient une perception tendant plutôt vers [Ø] dans 2 cas et plutôt vers [O] dans 3 cas[16]. Ainsi, 27% des réalisations de cette syllabe ont été perçues [O] par plus de 70% des auditeurs (44% par au moins 50% des auditeurs).

2.4. ANALYSE FORMANTIQUE

Une analyse formantique des réalisations montre que chez la plupart des locuteurs la perception /O/ est due à une valeur moins élevée du deuxième formant (figure 1).

Cependant, certaines perceptions postériorisées des /ə/ ne sont pas imputables à ce facteur acoustique. En particulier, chez les locuteurs masculins (figure 2), la valeur de F2 n'est pas sensiblement plus faible. Cependant, il est impossible ici d'effectuer une analyse statistique en terme de différence significative liée à des catégories (Mann-Witney par exemple) avec six valeurs par locuteur. De même il n'est pas envisageable de regrouper les valeurs de locuteurs différents même du même sexe. Nous ne pouvons donc, ici, parler que de tendances.

2.5. HYPOTHESES DE FACTEURS FAVORISANT LA POSTÉRIORISATION

Une analyse plus détaillée d'un corpus plus étoffé montrerait sans doute l'existence de facteurs linguistiques et extra-linguistiques favorisant les productions postériorisées.

2.5.1. FACTEURS LINGUISTIQUES

On peut tout d'abord supposer que l'harmonie vocalique puisse jouer un rôle certain dans un mot comme *recopiez*, mais tous les items postériorisés de notre corpus d'exemples ne présentent pas forcément un /O/ ni même une voyelle d'arrière en deuxième syllabe (cf. *requin* prononcé [rokɛ̃]). Dans le cadre d'un corpus ne concernant que deux mots, il n'est pas possible de tester la significativité de ce critère.

Un autre facteur possible serait celui de l'accentuation. Je pense en effet avec Fónagy (1979) ou Lucci (1983) que l'accentuation barytonique[17] est en cours de diffusion en français contemporain et en particulier chez les professionnels de la parole. Ainsi une syllabe initiale contenant un schwa verrait cette voyelle assurément prononcée et pourquoi pas colorée. J'ai calculé, pour chaque syllabe proposée aux auditeurs, un indice d'accentuation (I) en termes d'intensité (Ii) et de fréquence (If)[18]. La pertinence de cet indice n'est cependant pas statistiquement significative pour ce corpus ($p>0,06$). Autrement dit les *re* prononcés [rO] ne sont pas accentués différemment que ceux prononcés [rØ]. Ce qui ne veut pas pour autant dire qu'ils ne soient pas accentués.

2.5.2. FACTEURS EXTRA-LINGUISTIQUES

Tout d'abord, le facteur qui apparait comme déterminant dans cette étude est celui du sexe du locuteur: si 2 hommes sur 7 postériorisent leur /ə/ dans *recopiez*, le phénomène touche 6 femmes sur 9. Plusieurs chercheurs, dont Fònagy (1989), ont déjà noté que les femmes âgées de 20 à 35 ans produisent d'avantage de variantes "novatrices" que les hommes même du même âge. "Pour les changements phonétiques les femmes ont une génération d'avance sur les hommes" (Lennig 1979 : 33). Il semblerait que ce phénomène soit bien corrélé avec le caractère prestigieux de ces même variantes (Labov, 1976). Les locuteurs de ce corpus appartiennent à des classes d'âges différentes[19] mais toutes les femmes sont âgées de 20 à 32 ans. Les deux hommes producteurs de /ə/ postériorisés appartiennent aux classes A et B. Il est donc ici difficile d'isoler ce qui relève du facteur âge par rapport au facteur sexe.

Le corpus sur lequel j'ai fait mon étude n'a pas été établi à des fins sociolinguistiques. En particulier il ne

saurait en aucun cas être un échantillon représentatif de la population française. Tous les locuteurs sont soit étudiant, soit enseignant à l'université. Si dans l'ensemble, les étudiants sont plus producteurs de variantes postériorisées, il faut noter que ce corpus ne présente que des phrases lues, style plus formel qu'un cours ou on peut entendre de nombreuses variantes postériorisées chez les enseignants.

S'il n'est pas encore possible de le montrer ici, je pense que la postériorisation des /ə/ est essentiellement le fait des professionnels de la parole (professeurs et journalistes en premier lieu). Ainsi, ces variantes seraient plutôt prestigieuses et transmises par les jeunes femmes. Je suis en train de vérifier cette hypothèse.

Notes

* Quelle que soit la théorie phonologique adoptée, on peut être d'accord pour l'existence phonémique segmentale /ə/. On sait que sa réalisation varie entre [ø], [œ] et rien .

(1) Le terme "arrondies" comprend aussi les voyelles nasales /ɑ̃/ et /ɔ̃/. En ce qui concerne la postériorisation de /ɑ̃/ voir I. Malderez (1991)

(2) J'entends par *atypisme* une faute d'orthographe qui n'est pas prévue dans les exercices de régulation: si "des chevals" est une faute typique, "des chovaux" ne l'est pas.

(3) Les mots soulignés représentent les mots orthographiés.

(4) Je n'opère pas ici de distinction entre /ə/,/ø/ et /œ/ d'une part et /o/ et /ɔ/ d'autre part et traite chaque paire ou triplet comme un phonème unique.

(5) Ecoles J. de la Fontaine, Ully St Georges; G. Lagrené, Frocourt; J. Ferry, Catenoy.

(6) 'Complète avec e ou o'. Plouf le r....quin r....dait près d.... la plage. Il aperçut un r....ssignol qui r....muait sur le r....bord d'un r....cher. « Que fais-tu là ? lui demanda Plouf. — Je me r....pose mais je vais r....partir sinon je s....rai en r....tard au r....déo. — R....viens me raconter quand ça s....ra fini. Moi, je ne peux pas s....rtir de l'eau! — D'accord! A plus tard j....li r....quin. »

'Entoure le mot qui convient'. a- Pour me {peser, poser}, j'utilise une balance. b- J'ai buté dans une {motte, meute} de terre. c- L'écureuil est un animal très {poreux, peureux}. d- Il s'est {lové, levé} à 7 heures ce matin. e- Comme il fait froid, j'ai mis mon {bonnet, benêt}.

(7) 'Complète avec u ou ou'. Avec ma classe, je s_is parti deux j_rs en r_lotte. R_quin, notre cheval, s'est fait r_spéter car il préférait r_miner pl_tôt que d'avancer. On lui avait mis un joli r_ban r_ge autour du cou. Quand on le décrochait de la r_lotte, il galopait dans le pré et faisait des r_ades. Après il b_vait dans le r_isseau. La nuit, on a dormi dans des tentes et aussi on a j_é au r_gby dans un champ.

'Entoure le mot qui convient'. a- Le lion {rougit, rugit}. b- Quand je mange, je ferme la {bouche, bûche}. c- Pour écrire la maîtresse s'installe à son {bourreau, bureau}. d- Sans soleil les fruits ne peuvent pas {mûrir, mourir}. e- Les {poules, pulls} picorent dans la basse-cour.

(8) Dans les résultats qui vont suivre le lexème *motte* n'est pas pris en compte, je reviendrai plus loin sur ce cas particulier.

(9) Seuls ont été pris en compte les résultats des enfants ayant participé aux deux tests.

(10) liste A: t..lipe, s..oleil, p..qûre, m..man, r..blochon, p..ule, f..eur, r..clame, p..mme, c..ub. liste B: c..membert, g..uyère, t..me, c..èvre, r..blochon, b..ie, c..ntal, r..uy, f..urme, c..mté.

(11) Ce corpus n'a pas été créé pour les besoins de cette étude mais était disponible au laboratoire.
1- Il reviendra ce soir à cinq heures. 2- Le gardien a dit qu'il reviendra ce soir à cinq heures. 3- Le gardien a dit: "Il reviendra ce soir à cinq heures". 4- Dicte-moi après: "Il reviendra ce soir à cinq heures". 5- En entrant je vis son message: il reviendrait ce soir à cinq heures. 6- Recopiez le texte à partir de: "selon les pompiers, la bombe aurait été placée dans les cuisines.

(12) On peut reprocher à ce type de test un déséquilibre : le sujet s'attend à une équitable répartition des occurrences des réalisations proposées et peut ainsi être tenté de reconnaître plus de réalisations déviantes qu'un sujet à qui on aurait indiqué la voyelle attendue pour chaque item.

(13) A la plupart des locuteurs correspondent 7 items, parfois cinq - voyelle non prononcée - ou neuf -mot prononcé deux fois.

(14) Seuls 3 /ə/ sur 96 n'ont pas été prononcés (schwa).

(15) J'entends par "majoritairement" une perception à plus de 70 % des auditeurs.

(16) Ici les pourcentages de perception sont situé entre 50 et 70 %.

(17) Accentué ailleurs que sur la dernière syllabe, le plus souvent sur la première.

(18) Calculé comme suit : Ii ou If = (x-M)/s où x représente l'intensité ou la fréquence de la syllabe *re*; M, la moyenne des trois ou quatre syllabes qui composent les mots *recopiez* ou *il reviendra(it)* et s l'écart type de ces syllabes : I = Ii + If

(19) A : de 20 à 35 ans; B : de 40 à 55 ans

Références

CARTON, Fernand; ROSSI Mario; AUTESSERRE Denis & LEON Pierre (1983) *Les accents des français*, Paris: Hachette.

FONAGY, Ivan (1989) Le français change de visage ? *Revue romane, 24, 2*, 225-254.

---- & FONAGY, Peter (1971) [En hongrois : "Comment faire usage des fautes d'orthographe en hongrois ? "] in *Magyar Nyelvör, 95*, 70-89.

LABOV, William. (1976) *Sociolinguistique*, Paris: Seuil.

LENNIG, Matthew. (1979) Une étude quantitative du changement linguistique dans le système vocalique parisien. In P. Thibault (ed). *Le français parlé : études sociolinguistiques*, Edmonton : Linguistic Research, 29-39.

MALDEREZ, Isabelle. (1991) Neutralisation des voyelles nasales chez des enfants d'Ile-de-France. *Actes du XII ème Congrès International des Sciences Phonétiques, vol. 2*, Aix-en-Provence, 174-177.

MARTINET, André. (1969) C'est jeuli le Mareuc ! *Le français sans fard*, Paris: P.U.F., 191-208.

Tableaux

Tableau 4: Perception des /O/ attendus (locutrices puis locuteurs).

LOCUTEURS	Voyelle perçue en majorité	Pourcentage de perception [Ø]
ad	ø	57
bt	u	14
ch	o	14
dl1	ø	71
dl2	ø	71
evb	ø	57
hlgdb	o	29
kl	ø	57
xx	ø	50
xxx	ø	43
bl	ø	57
gb	ø	71
gg	ø	50
jydo	ø	93
jydu1	ø	50
jydu2	ø	71
kh	ø	64
ls	ø	79
	moyenne	55

Tableau 3: test adultes: r...blochon

	liste A		liste B		Total	
	e	o	e	o	e	o
répartition	1	5	3	3	4	8
% répartition	17	83	50	50	33	67

Tableau 2: Pourcentage d'erreurs par rapport au nombre de réponses exprimées pour chaque item et pour chaque cours (Moyennes 1: *motte* inclus, moyennes 2: *motte* exclu).

	CE1	CE2	CM1	CM2
requin 1	39,29	28,85	23,08	2,99
rodait	22,92	17,65	6,00	2,99
de	0,00	3,77	0,00	0,00
rossignol	23,64	22,64	9,80	0,00
remuait	21,43	20,75	7,69	0,00
rebord	17,86	7,69	5,77	0,00
rocher	17,24	11,32	9,62	2,99
repose	15,52	5,66	3,85	2,99
repartir	14,04	7,55	0,00	1,49
serai	15,79	7,55	1,92	1,49
retard	7,02	5,66	0,00	0,00
rodéo	30,77	15,09	5,77	1,49
reviens	8,93	1,96	0,00	0,00
sera	10,71	5,66	1,92	0,00
sortir	8,93	1,92	0,00	0,00
joli	0,00	9,43	1,92	0,00
requin 2	43,86	24,53	21,15	2,99
peser	14,29	15,09	13,73	2,99
motte	36,00	41,51	60,78	50,75
peureux	16,07	0,00	5,88	0,00
levé	7,14	1,89	0,00	0,00
bonnet	8,77	0,00	3,92	0,00
moyennes 1	17,88	11,66	8,27	3,32
moyennes 2	16,39	10,22	5,81	1,07

Tableau 1: Pourcentages d'erreur au test en fonction des contextes, des voyelles et des niveaux

Contextes		V	CE1	CE2	CM1	CM2	Pop totale
/r/ — C_1^2 V (C) (V)	(rodait, rossignol, ...)	/O/	23,47	16,67	7,80	1,87	11,83
/r/ — C_1^2 V (C) (V)	(requin, reviens, ...)	/ə/	20,97	12,86	7,69	1,31	10,30
C[-r] — C_1^2 V (C) (V)	(joli, bonnet, ...)	/O/	5,85	3,80	1,94	0,00	2,77
C[-r] — C_1^2 V (C) (V)	(serai, sera, ...)	/ə/	10,65	5,68	3,88	0,75	5,05
C — C_1^2 V (C) (V)	C initiale indifférente	/O/	15,63	11,14	5,28	1,07	7,91
C — C_1^2 V (C) (V)	C initiale indifférente	/ə/	16,56	9,77	6,07	1,07	8,05
	résultats globaux		16,39	10,23	5,81	1,07	8,01

Figures: Représentation dans le plan F2/F1 des réalisations phonétiques [O] des /ə/ de *recopiez* et [Ø] des /ə/ de *reviendra(it)* ainsi que d'un /i/ et d'un /a/:

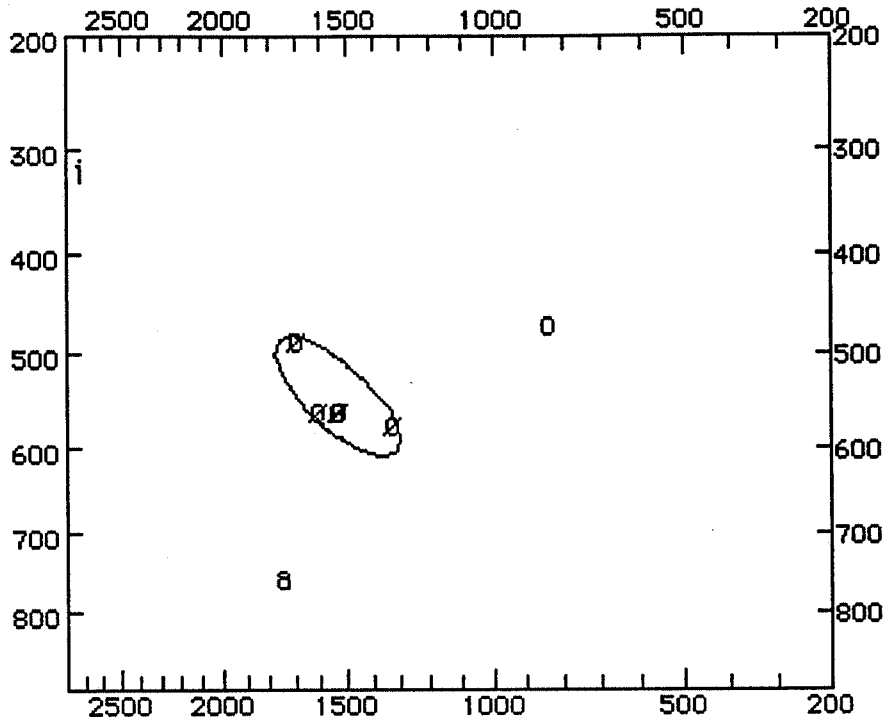


Figure 1 : locutrice CH

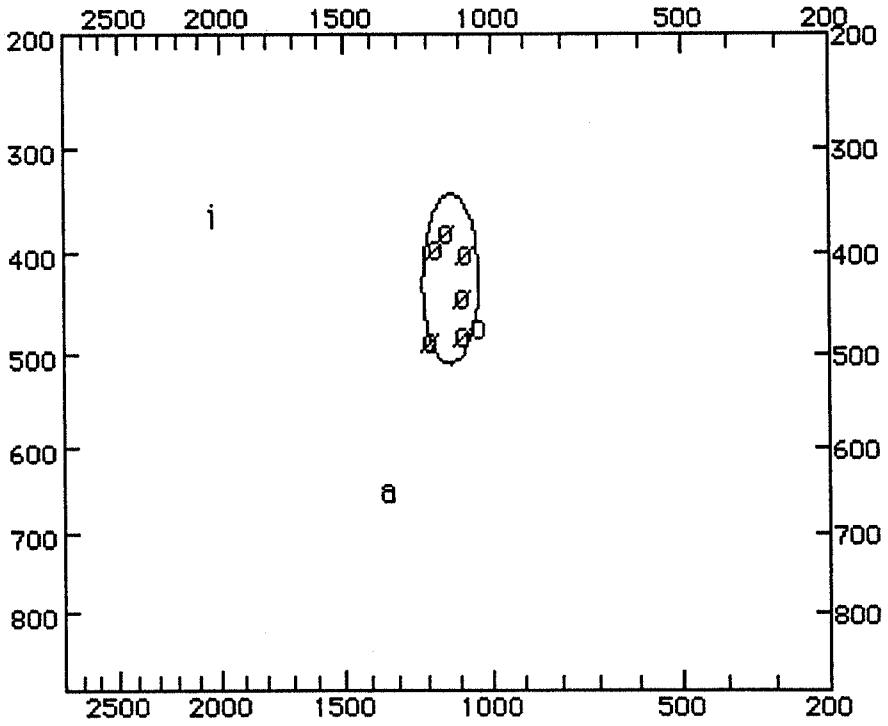


Figure 2 : locuteur JYDU

FAITS SYSTÉMATIQUES DE DURÉE EN FRANÇAIS QUÉBÉCOIS, RYTHME ET ACCENTUATION.

MARISE OUELLET

CIRAL, UNIVERSITÉ LAVAL, QUÉBEC

Résumé

Une étude sur les modalités d'agencement des durées segmentales dans les syllabes en français québécois a permis de mettre en évidence les liens étroits qui existent au sein des rimes syllabiques. Nous verrons que cette cohésion ..vc$, d'abord observée en syllabes fermées s'étend à d'autres contextes, notamment en syllabes dites ouvertes et qu'elle peut être associée à certaines manifestations particulières de l'accent.

mots-clefs: durée segmentale, syllabe, accent, rythme.

PROBLÉMATIQUE

Le système vocalique du français québécois est constitué de dix-sept voyelles dont huit sont longues: orales longues / ɑ ø ɜ o/ et voyelles nasales /ɑ̃ ɛ̃ õ œ̃/. Ces voyelles longues partagent un certain nombre de traits communs sur la base desquels Santerre (1987) les regroupe au sein d'une même classe, celle des voyelles longues par nature, qu'il oppose aux neuf voyelles restantes qualifées de brèves. Outre des différences systématiques de leurs durées moyennes, les voyelles longues se distinguent des voyelles brèves au plan de la variabilité sous l'effet de l'entourage consonantique immédiat. Ainsi, la durée d'une voyelle brève allongée peut atteindre celle d'une voyelle longue orale voire même celle d'une voyelle nasale, ce qui laisse supposer que la durée absolue ne constitue pas l'élément sur lequel repose la distinction brève-longue. Si l'on s'en tient aux voyelles orales longues ayant des contreparties brèves (a/ɑ ɛ/ɜ œ/ø ɔ/o) on constate que les voyelles sont aussi différenciées par le timbre. Les voyelles longues (incluant les nasales) peuvent être diphtonguées et le sont en général, alors que les voyelles brèves ne seront diphtonguées qu'en contexte consonantique allongeant, ce qui peut toutefois amoindrir l'impact des variations du timbre comme indice d'identification des voyelles orales longues et brèves. Il en ressort que les oppositions entre voyelles longues et brèves ne reposent pas entièrement sur des contrastes de timbre ni sur des contrastes systématiques entre les durées absolues.

En français québécois, les distinctions entre voyelles brèves et voyelles longues sont attestées en syllabe libre et entravée à l'intérieur du mot ainsi qu'en syllabe entravée finale de mot (Gendron 1966, Boudreault 1968). Un lien a été rapidement établi par Gendron (1966) entre la persistence des voyelles longues et leur confinement à certaines bases lexicales (/ɜ/ présent dans "fête" et dans toutes les formes suffixées "fêtard", fêter" etc.). Santerre (1987) associe également la présence des voyelles longues à certains morphèmes lexicaux et mentionne que la coupe morphologique aurait priorité sur la coupe syllabique canonique en français québécois de telle sorte que la consonne /t/ de "*pâte*", dans "*pâte à tarte*" ne formerait pas d'enchaînement avec la voyelle subséquente.

Ce système est différent de celui encore d'usage au siècle dernier dans l'Hexagone, lequel a progressivement disparu. Cette disparition progressive des oppositions a/ɑ et ɛ/ɜ est perceptible dans les travaux de phonétique par les choix des unités sous lesquelles seront regroupées les valeurs de durée des segments (deux "A" et deux "E" chez Delattre 1959, deux "A" chez Brichler-Labaye 1970, /a/ et /ɛ/ chez Di Cristo 1985, Bartkova et Sorin 1987). Il est toutefois possible de rencontrer certains surplus de durée accompagnant la production des anciennes voyelles longues dans le français commun (Gottfried et Beddor 1988), hors des parlers dialectaux où des oppositions systématiques de durée sont encore attestées (Cheik

Rouhou et al. 1989). Nous verrons que ces surplus de durée n'ont pas d'incidence sur la structure temporelle des syllabes, mais que, par contre, en français québécois, les oppositions phonologiques sont maintenues par le biais de modalités particulières d'agencement des durées segmentales au sein de la syllabe.

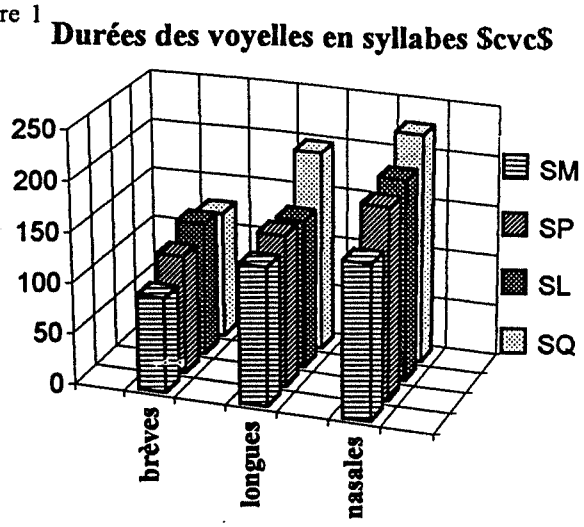
Dans cet exposé, nous procéderons d'abord à la description des modalités d'agencement des durées segmentales en syllabes accentuées cvc. Nous verrons ensuite que le problème posé par la division syllabique des morphèmes lexicaux découle d'une propriété plus générale des syllabes en français québécois. Enfin, nous montrerons, en ayant recours aux résultats obtenus dans diverses études, comment le lien étroit existant entre les constituants de la rime syllabique semble conditionner l'accentuation en franco-québécois.

SYSTÉMATIQUE DES DURÉES

Afin d'établir un modèle de l'organisation temporelle basé sur les valeurs des durées segmentales et les modifications qu'elles subissent à l'intérieur de la syllabe, nous avons effectué un premier ensemble d'observations en syllabes fermées CVC, sous l'accent. Le modèle dégagé permet d'établir une certaine hiérarchie entre les constituants de la syllabe (Ouellet 1992), hiérarchie qui suppose l'existence d'une cohésion plus grande entre les segments ..vc$ en français québécois et entre les segments $cv.. en français de l'Hexagone. Nous limiterons ici nos commentaires aux voyelles brèves /a ɛ/, longues /ɑ ɔ/ et nasales /ɑ̃ ɛ̃/. Les exemples présentés dans la première et la seconde partie de cet exposé sont des extraits d'un corpus plus vaste constitué de séquences monosyllabiques et disyllabiques. Les résultats présentés ci-après sont tirées de syllabes et de mots cibles insérés dans des phrases porteuses ("Le mot ... me plaît") soumises pour lecture à deux locuteurs québécois (un Montréalais (SM) et un locuteur de la région de Québec (SQ)) et à deux locuteurs français (l'un Parisien (SP) et l'autre Lyonnais (SL)). Nous nous concentrerons ultérieurement sur les données des corpus rassemblés à Montréal et à Paris à partir desquels le modèle syllabique proposé a été établi.

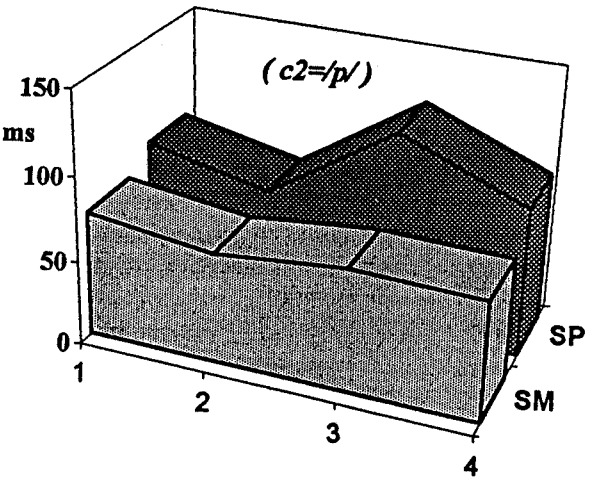
Voici les valeurs obtenues pour les voyelles brèves (ex: patte, faites), les voyelles longues orales (ex: pâte, fête) et les voyelles nasales (ex: pente, feinte) en syllabes cvc prononcées par deux locuteurs québécois (SM et SQ) et deux locuteurs français (SP et SL). On notera que chez les deux sujets français, un supplément de durée accompagne la production des anciennes voyelles longues /ɑ/ et /ɔ/ . Chez les sujets québécois, les valeurs de durée de ces mêmes voyelles longues se rapprochent sensiblement des valeurs obtenues pour les voyelles nasales.

figure 1



Dans la syllabe, la dynamique d'ajustement des durées n'est toutefois pas la même selon qu'il s'agit du locuteur français ou du locuteur québécois. Si l'on se base sur les variations co-intrinsèques de durée, on constate que les consonnes initiales n'affectent pratiquement pas les noyaux vocaliques en français québécois alors qu'elles entraînent des variations de durées pour les voyelles chez SP. Dans la figure qui suit, nous présentons les variations des durées du /a/ bref après des consonnes sourdes en 1 et 2 et des consonnes voisées en 3 et 4.

figure 2



Chez les deux informateurs, la consonne post-vocalique exerce une grande influence sur les valeurs des durées des noyaux syllabiques. Ce qui différencie principalement nos locuteurs au plan de la gestion des durées c'est le positionnement du lien plus étroit entre les constituants de la syllabe. Ce lien est observable par le biais des variations co-intrinsèques et compensatoires des durées segmentales. Chez notre sujet québécois, toute modification de la durée intrinsèque du noyau vocalique implique un ajustement de la durée de la coda

($\rho<0.0001$) ce qui n'est guère la cas lorsque l'on met en rapport les durées des voyelles avec celles des consonnes qui lesprécèdent ($\rho< 0.5888$). Ce même type de rapport compensatoire est relevé entre l'attaque consonantique et le noyau vocalique chez notre informateur parisien ($\rho<0.0001$). Si les durées intrinsèques des voyelles constituent la principale source de modification des durées des codas, la durée de la consonne initiale de syllabe en français québécois ne réagit qu'à la durée totale de la rime. Les figures présentées ci-après illustrent la variabilité de l'attaque consonantique /p/ en français québécois par rapport aux modifications de durée de la voyelle sous l'effet de différentes consonnes postposées, sourdes sonores, occlusives et constrictives. Dans la seconde figure, on constate l'invariabilité de l'attaque consonantique chez notre locuteur français.

figure.3

Incidence de la rime sur la durée de l'attaque consonantique

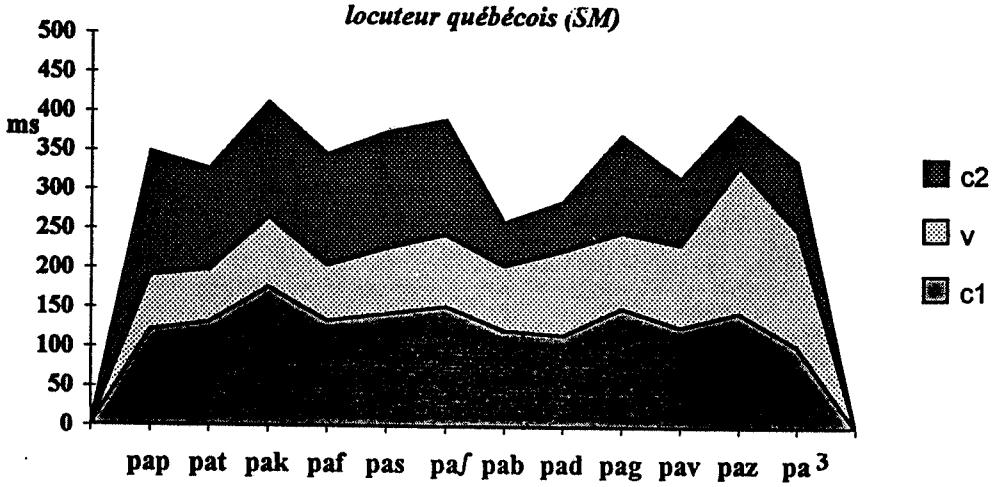
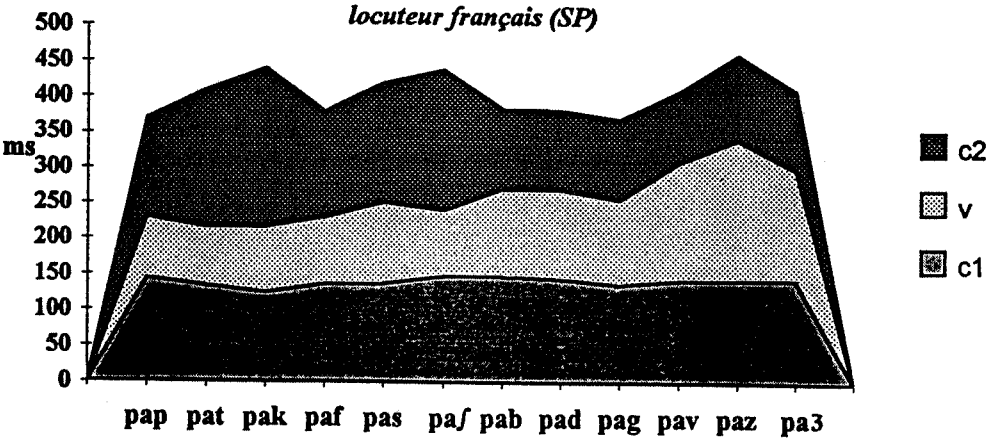


figure 4

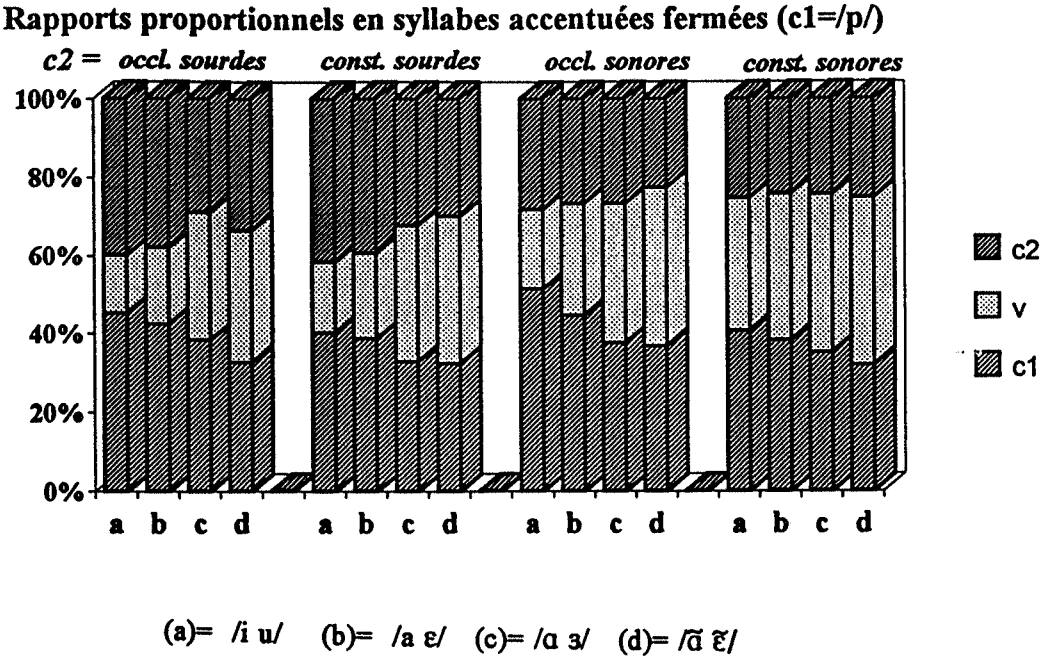
Incidence de la rime sur la durée de l'attaque consonantique



Cette dynamique ne prend de sens que si elle est envisagée dans une perspective où la répartition des durées segmentales dans la syllabe contribue à mettre en relief les propriétés phonologiques des segments. La syllabe et la rime semblent constituer les terrains priviligiés où s'actualisent les oppositions de durée en français québécois. Ainsi, lorsque des voyelles longues et brèves deviennent accidentellement isochrones, leur identité est préservée par un jeu de proportions à l'intérieur de la syllabe (durée du segment en % sur la durée totale de la syllabe).

Nous avons constaté que l'identité d'un segment vocalique ou consonantique est plus facile à établir lorsqu'on observe la portion de la syllabe qui lui est dévolue plutôt que sa durée objective. La figure qui suit montre un exemple de répartition de "l'espace" syllabique avec des voyelles de durée intrinsèque différente, soit /i/ et /u/ en (a), les brèves /a/ et /ɛ/ en (b), les voyelles longues /ɑ/ et /ɜ/ en (c) et les voyelles nasales /$\tilde{\alpha}$/ et /$\tilde{\varepsilon}$/ en (d). Les attaques consonantiques sont constituées par la consonne /p/ alors que les codas sont constituées d'occlusives sourdes ou sonores et de constrictives sourdes ou sonores.

figure 5



(a)= /i u/ (b)= /a ɛ/ (c)= /ɑ ɜ/ (d)= /ɑ̃ ɛ̃/

L'observation des agencements des durées dans les syllabes entravées en français québécois nous a conduit à proposer un modèle dynamique de l'organisation temporelle traduisant les liens étroits entre les segments formant la rime (C1 <-- / VC2), ce qui diffère du modèle que nous avons obtenu par l'analyse des syllabes prononcées par un sujet parisien, où cette même cohésion se manifestait entre l'attaque et le noyau subséquent (C1 V/ --> C2). Des résultats similaires ont été obtenus par Rostolland et al. (1985) avec des sujets français.

PROPRIÉTÉS DE LA RIME ET PRINCIPES DE SYLLABATION

Nous avons pu constater que les modalités d'agencements des durées qui prévalent dans les syllabes cvc en français québécois contribuent à mettre en relief les propriétés temporelles de tous les segments; voyelles longues, brèves, consonnes sourdes, sonores, etc. Le lien étroit qui a été dégagé entre les constituants de la rime syllabique peut-il être à l'origine des phénomènes observés par Santerre (1987) et qui ont trait à la conservation des consonnes finales dans les morphèmes lexicaux?

Dans le but de répondre à cette question, nous avons procédé à des mesures de durée de segments cibles, d'une part, dans des mots de deux syllabes constitués de bases suffixées et, d'autre part, dans des mots de deux syllabes ne comportant pas de telles bases (par exemple /pɛ̃s/ dans (a) "*pince*", (b) "*pincement*", (c) "*pincer*" puis dans (d) "*pinson*"). Les mots porteurs des séquences cibles pouvaient à priori être syllabés comme suit: *#cvc#* en (a), *#cvc$cv#* en (b) et *#cv$cv#* en (c) et (d). Partant des mesures de durées obtenues pour c1, v et c2 en syllabes accentuées (a), nous nous sommes attardées aux phénomènes de compression et d'expansion des durées en tant qu'indices d'un déplacement des frontières syllabiques. Le passage du contexte (a) au contexte (b) représente un éloignement par rapport à l'accent. De (b) à (c), la séquence cible passe de l'état de syllabe entravée à celui de syllabe libre, bien que les trois segments fassent toutefois partie d'une base ou d'un morphème lexical. Enfin, du contexte (c) au contexte (d), le lien morphologique unissant théoriquement les trois segments cibles est rompu.

Les résultats obtenus ne nous ont pas permis de différencier de façon systématique le comportement des voyelles brèves et des voyelles longues. Ainsi, sans égard à la nature brève ou longue de la voyelle, les transformations observées quant aux durées des segments cibles c1, v et c2 sont sensiblement les mêmes et se résument aux faits suivants: du contexte (a) au contexte (b) la consonne initiale et la voyelle sont abrégées; au passage de (b) à (c), la voyelle, mais particulièrement la consonne c2 sont allongées alors que c1 demeure stable; au passage de (c) à (d) la consonne initiale c1 et la consonne finale connaissent une expansion de leurs durées respectives assortie d'une faible augmentation de la durée de la voyelle. Les moyennes de durée établies dans les quatre contextes à l'étude pour chacune des segments cibles c1, v et c2 sont présentés dans le tableau qui suit.

tableau 1

Durées des segments cibles
(tous segments confondus)
les durées sont exprimées en ms

	Sujet québécois
	c1 v c2
(a)	124 113 94
(b)	105 90 95
(c)	103 94 114
(d)	112 97 123

n=39 segments par contexte

Si l'on admet avec, entre autres, Kozhevnikov et Chistovich (1965) que plus une syllabe renferme de segments plus les durées de ces segments seront comprimées, nos résultats sembleraient plutôt appuyer l'hypothèse d'un enchaînement de la consonne c2 avec la voyelle suivante (ou déplacement de la frontière syllabique) au passage de (b) à (c). Cette consonne, dans la mesure où elle forme un enchaînement avec la voyelle suivante, devient l'attaque de la syllabe finale, ce qui pourrait expliquer en partie son allongement. Toutefois, la faible expansion des durées vocaliques et la stabilité des durées des consonnes initiales passant de (b) à (c) laissent plutôt croire que la syllabe initiale demeure entravée (contrairement à ce que l'on observe de (c) à (d)). L'hypothèse qu'il est alors possible d'émettre est que la consonne faisant partie d'un morphème lexical conserve son statut de coda à l'intérieur du morphème mais joue également le rôle d'attaque consonantique dans la syllabe finale, ce qui lui confèrerait un statut de consonne ambisyllabique. Le cas échéant, le modèle dégagé dans les syllabes fermées sous l'accent et dont la dynamique contribue à la mise en relief des propriétés temporelles des segments dont les propriétés phonologiques des voyelles pourrait être étendu à d'autres contextes.

Mentionnons enfin que le corpus initial duquel nous avons extrait ces données comportait deux séries d'enregistrements, soit une série où la phrase porteuse était " Le mot ... me plaît" et une seconde série dont les résultats ne sont pas présentés ici et pour laquelle la phrase porteuse était " Le mot ... sonne bien". Les faibles transformations rythmiques induites par le changement dans la structure des phrases porteuses (ici, la disparition du pronom complément devant le verbe de la première à la seconde série) ont systématiquement affecté les durées des rimes syllabiques des mots cvc chez notre sujet québécois. Un phénomène similaire a été observé par Cedergren et *al.* en français québécois spontané. Chez notre informateur français, ces changements rythmiques ont modifié les durées de la consonne initiale et de la voyelle subséquente.

DISCUSSION

Certains rapprochements peuvent être effectués entre différents phénomènes reliés à la durée segmentale, l'accentuation et le rythme en français canadien et les observations qui viennent d'être faites en ce qui a trait au statut de la consonne post-vocalique et à la structure temporelle des syllabes en général. La tendance à priviligier la réalisation de syllabe de type cvc ou vc$ nous semble constituer l'une des caractéristiques du français parlé au Québec.

L'un des phénomènes observés régulièrement dans les études sur le rythme ou l'accentuation en français québécois a trait à l'allongement des syllabes pénultièmes. Dans sa description du rythme et de l'intonation du français parlé au Canada et en France, Boudreault (1968) mentionne l'existence de syllabes pénultièmes dont les durées laissaient supposer la présence d'un accent. L'auteur mentionne que ces syllabes sont soit entravées, soit longues par la voyelle; mais il retrouve dans certaines syllabes longues des voyelles brèves, notamment le /ɛ/. Bien que l'allongement des syllabes pénultièmes se produise aussi dans la variété du français de France qu'il a étudiée, Boudreault considère que l'allongement des syllabes prétoniques constitue un phénomène caractérisant les prononciations de ses sujets canadiens, et qu'il est probable que ces syllabes longues ne soient perçues comme telles que par une oreille étrangère. Toutefois, dans une étude sur la perception de l'accent en français québécois, Paradis et Deshaies (1990) ont pu constater qu'un accent était régulièrement détecté par les auditeurs sur les syllabes pénultièmes. Les auteurs ont démontré que l'on pouvait augmenter le pouvoir de prédiction d'un modèle d'analyse métrique de l'énoncé en tenant compte du poids des syllabes et en conférant à certaines consonnes intervocaliques le statut de consonnes ambisyllabiques.

La présence de ces syllabes pénultièes longues ne semble toutefois pas entraver la démarcation de fin de mot et de syntagme. Une récente étude effectuée par Santerre (1991) a mis en évidence un phénomène particulièrement intéressant et qui a trait à la réalisation d'accents adjacents en français québécois. Selon Santerre, la fin du syntagme et du mot serait marquée par l'intonation plutôt que par la durée. La présence de syllabes longues en pénultième, jumelée à la présence d'une marque intonative (le plus souvent une montée) en syllabe finale constituent, de l'avis de l'auteur, deux marques accentuelles adjacentes qui ne s'excluent pas mutuellement, ce qui va à l'encontre des principes

d'accentuation du français proposés par Dell (1984).

Le lien que nous avons dégagé entre les constituants de la rime syllabique a également été observé en français québécois spontané, en particulier les ajustements des durées des voyelles et des codas. (Cedergren et *al.*, 1993). Le rythme d'une langue étant modelé par la forme canonique des syllabes (Vaissière, 1991), les propriétés temporelles des segments et les modalités d'assignation de l'accent (Dauer, 1983), il appert que le français québécois a subi certaines mutations par rapport, par exemple, au français de l'Hexagone, notamment en ce qui a trait à la division syllabique. En français québécois, il semble que la dynamique qui prévaut à l'intérieur de la syllabe entrave dans une certaine mesure l'enchaînement des consonnes avec les voyelles suivantes, ce qui réduit la prépondérance des syllabes ouvertes et, d'un point de vue impressionniste, accentue les différences entre les variétés française et québécoise du français. Par ailleurs, il semble qu'il faille proposer des règles d'assignation de l'accent susceptibles de tenir compte du poids syllabique. La présence d'oppositions entre voyelles brèves et voyelles longues n'est sans doute pas étrangère à toutes ces mutations puisqu'elles ne s'actualisent de façon systématique qu'à travers la syllabe par un jeu des proportions qui a de toute évidence la syllabe comme domaine.

Il semble donc que pour comprendre l'interaction qui existe entre contraintes de bas et de haut niveaux dans la prosodie du français québécois, il faille prendre en compte l'une des caractéristiques les plus importantes de cette variété, soit la présence d'oppositions phonologiques de quantité à l'intérieur du système vocalique.

bibliographie

Bartkova, K. et C. Sorin, (1987) "A Model for segmental duration for speech synthesis in French", *Speech Communication*, vol. 6, pp 245-260.

Boudreault, M, (1968) Rythme et mélodie de la phrase parlée en France et au Québec, Paris Klincksieck et Québec, P.U.L., 273 p.

Brichler-Labaye, C. (1970) Les voyelles françaises, Paris, Klincksieck.

Cedergren, H., Cichocki, W., Perreault, H. et L. Levac (1993)
"Modeling prosodic effects on syllable duration in spontaneous speech", communication présentée au *Congrès Annuel de l'Association Canadienne de Linguistique*, Congrès des Sociétés Savantes, Ottawa, mai 1993.

Cheik Rouhou, M., Abry, C., et L.-J. Boe (1989)"À quelques kilomètres d'Hautreville... et près d'un demi-siècle après Martinet: premières mesures de la quantité vocalique en francoprovencal de St-Pierre-de-Soucy", *Bulletin de l'Institut de Phonétique de Grenoble*, vol.17-18, pp 249-270

Dauer, R.M. (1983) "Stress-timing and syllable timing reanalyzed", *Journal of Phonetics*, vol. 11, pp 51-62.

Delattre, P. ,(1959) "Rapports entre la durée, le timbre et la structure syllabique en français", *The French Review*, vol. 22, no 6, pp. 547-552.

Di Cristo, A. (1985) De la micromélodie à l'intono-syntaxe, Publications de l'Université de Provence, Aix-en-Provence, Tomes I et II.

Dell, F. (1984) «L'accentuation dans les phrases françaises» dans Formes sonores du langage, Dell *et al.* Eds, Paris, Hermann, pp.65-122

Gendron, J.-D. (1966) Tendances phonétiques du français parlé au Canada, Paris Klincksieck Québec, P.U.L., 254 p.

Gottfried, T.L. et P.S. Beddor, (1988) «Perception on temporal and spectral informations in French vowels», *Language and Speech*, vol. 31, part 1, pp 57-75.

Kozhevnikov, A. et L.A. Chistovich (1965), «Speech: articulation and perception», *U.S. Dept. of Com. Tech. Inf. Serv.* JDRS 30543

Ouellet, M. (1992) Systématique des durées segmentales dans les syllabes en français du Québec et de France, Thèse de doctorat, Université de Montréal, Montréal, 413p.

Paradis, Cl. et D. Deshaies (1990) «Rules of stress assignment in Quebec French; Evidence from perceptual data», *Language Variation and Change*, vol. 2 , pp 135-154.

Santerre L. (1991) «Incidences du trait phonologique de durée vocalique sur la prosodie en français québécois», *Actes du XII Congrès des Sciences Phonétiques*, Aix-en-Provence, vol. 4, pp 254-257.

(1987) «Systématique des durées segmentales dans les rimes syllabiques à voyelles longues et brèves par nature», *Actes du Congrès des Sciendes Phonétiques*, Tallinn, vol. 5 pp 120-129.

Vaissière, J. (1991) «Rhythm, Accentuation and Final Lengthening in French», *Music, Language, Brain and Speech*, R, Carlson, L. Nora et J. Sunberg Dir. MacMillan Press, pp. 188-200

LA PHONÉTISATION DU FRANÇAIS : LOGIQUE ET LINGUISTIQUE DANS UNE GRAMMAIRE-LEXIQUES

Rabia BELRHALI, Véronique AUBERGÉ, Louis-Jean BOË

INSTITUT DE LA COMMUNICATION PARLÉE
CNRS URA 368, INPG/Univ. Stendhal, BP 25 38040 Grenoble cedex 9
belrhali@icp.grenet.fr, auberge@icp.grenet.fr, boe@icp.grenet.fr

Résumé
Cette étude sur les règles de phonétisation du français a été menée sur la base de l'exploration systématique d'un grand lexique de formes canoniques phonétiques de référence. La grammaire-lexiques, écrite en langage TOPH, qui résulte de cette analyse met clairement en évidence les phénomènes linguistiques, en particulier étymologiques, mis en cause dans la transcription phonétique.

Mots clés : *phonétisation, grammaire-lexiques, organisation linguistique, étymologie, le langage TOPH.*

I. INTRODUCTION

Notre objectif est de proposer une modélisation linguistique des relations entretenues par la forme orthographique et la forme phonétique d'un message. Dans cette première étape, nous avons centré notre étude sur la phonétisation des mots. Les phénomènes de jonction seront modélisés sur la base de la grammaire des mots. Le principe méthodologique de cette étude est l'exploration systématique d'un lexique du français *Le 60 000* [Belrhali 92] qui a engendré la grammaire de phonétisation. Nous exposons d'abord la méthodologie adoptée pour définir toutes les correspondances graphème(s)-son(s). Ensuite nous montrons que les informations linguistiques véhiculées à l'oral dans ces relations de phonétisation sont distribuées selon une hiérarchie qui s'établit inclusivement sur les niveaux du graphème, du morphophonème, de l'unité lexicale, de la catégorie, de la syntaxe et de la sémantique. Nous voulons montrer que la grammaire logiquement établie par cette méthodologie de recherche et vérification systématique sur le lexique répond à une « lecture » linguistique. Ainsi nous décrivons comment les grandes classes d'irrégularités ont des comportements linguistiques réguliers au vu de leur étymologie. Nous montrons enfin comment les solutions proposées par la grammaire pour les cas non ambigus sont modélisables : l'architecture du modèle de formalisation adopté TOPH [Aubergé 87] impose à la grammaire une représentation suffisante pour la plupart des cas ; celle des règles activant des lexiques.

II. MÉTHODOLOGIE

TOPH (Transcription Orthographique-Phonétique) est un langage et un environnement logiciel développé à l'Institut de la Communication Parlée pour la synthèse de la parole [Aubergé 87]. Ce phonétiseur propose une syntaxe de description des grammaires de phonétisation sur texte libre. Il permet de réécrire une chaîne d'entrée graphémique en une chaîne de sortie phonétique. La formalisation des règles est transparente à l'utilisateur, permettant ainsi des modifications aisées. L'expert formalise son raisonnement sous la forme d'une grammaire déterministe de règles de réécriture contextuelles.
La grammaire comprend :
• des ensembles prédéfinis de caractères orthographiques décrivant des phénomènes de nature très différente :
- ensembles linguistiques :
« Consonnes nasales » = (m, n)
- ensembles d'exceptions :
« lexique des mots dont -er final se prononce [œʀ] » = (aft, angledoz, bookmak, boost, bootlegg, bulldoz, cantilev, clipp, cruis, dispatch, drift, dump, farm, freez, führ, jigg, kipp, lead, load, mak, mix, outrigg, outsid, ov, partn, pow, quaqu, rac, retriever, sell, sitt, spinnack, sweat, tank, thrill, trott, walkov, woof)
Ce lexique concerne les mots suivants :
after-shave, angledozer, bookmaker, booster, bootlegger, bulldozer, cantilever, clipper, cruiser, dispatcher, drifter, dumper, gentleman farmer, freezer, führer, hurdler, jigger, kipper, kniker, leader, loader, pace-maker, mixer, outrigger, outsider, crossing-over, partner, horse-power, racer, retriever, best-seller, spinnaker, sweater, triller, walkover, woofer
• des commentaires pouvant être une chaîne quelconque bornée par '!' et insérée dans n'importe quelle portion de la grammaire.
règle : (od) + en + (dron) = [ɛ̃]
commentaires : ! cette règle concerne les mots tel que lépidodendron, philodendron, mais pas Tendron !
• des règles ordonnées en classes; la classe d'une règle étant déterminée par le premier caractère de la chaîne à transcrire.

CLASSE DU A

Règles...

CLASSE DU B

Règles...

•

•

•

CLASSE DU Y

Règles...

CLASSE DU Z

Règles...

Les règles sont regroupées en 26 classes, chaque classe correspondant au cardinal de l'alphabet. L'agencement des règles les unes par rapport aux autres joue un rôle primordial et constitue un outil logique pour la formalisation linguistique des règles. En effet lors de la transcription, les règles sont examinées dans l'ordre où elles sont écrites, la première règle validée est appliquée sur le texte d'entrée. Les contextes gauche et droit de la chaîne à transcrire pondèrent sa réécriture.

Afin de tester et compléter la grammaire de phonétisation existante, un lexique de grande taille était nécessaire. Pour cela nous avons utilisé *Le 60 000* de l'ICP implanté sur Macintosh dans un environnement HyperCard. Afin de tester nos règles et d'obtenir la trace des règles appliquées, nous avons mis au point un environnement convivial autour de TOPH : Le Phonétiseur.

L'environnement HyperCard et le langage HyperTalk ont rendu possible la mise au point de programmes de recherche de chaînes de caractères orthographiques de longueur quelconque. Elles ont été recherchées dans trois positions : initiale, interne, et finale de mot. À partir des listes obtenues nous avons systématiquement relevé la transcription phonétique de la chaîne étudiée en prenant comme référence de prononciation le Petit Robert 1. Nous avons ensuite vérifié l'existence de la (ou des) règle(s) correspondante(s) à la (ou aux) transcription(s) phonétique(s) de la chaîne de caractères étudiée. Dans le cas contraire, nous avons écrit de nouvelles règles.

Exemple : la classe du « B »

Nous avons obtenu 2 394 mots commençant par "b-", 7 210 mots contenant au moins un "-b-" en position interne et 32 mots se terminant par "-b". Après le relevé de la prononciation du graphème "b" dans tous les mots et dans toutes les positions, nous avons établi les règles suivantes :

- ("#") + b + ("#") = (be)

Cette règle concerne la lettre de l'alphabet, et peut être utilisée dans des sigles.

- + b + = (b)

Il s'agit de la règle la plus générale.

- + b + (s, t) = (p)

Cette règle concerne tous les mots contenant la suite "bs" ou "bt" et dont le "b" se réalise [p]; il s'agit ici d'un cas d'assimilation régressive.

- (su) + b + (sidence, sidiare, sist) = (b)
- (bo, lam) + b + (sleigh, swool) = (b)

Ces deux dernières règles sont des exceptions à la précédente.

- + b + (c, k) = (p)

Cette règle concerne tous les mots contenant la suite "bc" ou "bk" dont le "b" se réalise [p]. Nous avons ici un autre cas d'assimilation régressive. La seule exception à cette règle est la suivante :

- (su) + b + (carpatique) = (b)
- + bb + = (b)

Toutes les géminées de la classe du "b" obéissent à une règle unique.

- (radou, lom) + b + ("#", s) = ( )

Cette règle concerne radoub et les mots se terminant par "lomb" comme plomb, coulomb, surplomb, aplomb, dont la réalisation du "-b" en position finale est muette (ces mots peuvent être éventuellement suivis d'un "s", marque du pluriel).

Classement suivant l'ordre d'application des règles :

- (radou, lom) + b + ("#", s) = ( )
- (bo, lam) + b + (swool, sleigh) = (b)
- (su) + b + (sidence, sidiare, sist) = (b)
- + b + (s, t) = (p)
- (su) + b + (carpatique) = (b)
- + b + (c, k) = (p)
- + bb + = (b)
- ("#") + b + ("#") = (be)
- + b + = (b)

Cette méthodologie a été appliquée pour toutes les lettres de l'alphabet afin de cerner le plus exhaustivement possible toutes les différentes réalisations phonétiques d'un graphème.

III. LES DIFFÉRENTS NIVEAUX LINGUISTIQUE

3.1 Le niveau graphémique

La phonétisation véhicule des informations linguistiques de différents niveaux : graphémique, morpho-phonémique, lexical, catégoriel, syntaxique et sémantique.

Le premier niveau correspond à la réécriture régulière d'une chaîne graphémique en une chaîne phonétique. C'est donc au niveau phonotactique où les règles ne tiennent pas compte de la forme globale du mot et *a fortiori* de niveaux plus complexes. Par exemple :

- la chaîne "eau" se transcrit [o]
- si "s" est entre deux voyelles se réécrit [z]
- sinon "s" se réécrit [s]

3.2 Le niveau lexical

Un ensemble de mots peuvent partager une même particularité de réécriture, bien souvent expliquée par des faits semblables (mots d'emprunt de même origine, de même usage...). Ces ensembles de mots seront réunis dans des lexiques. Il arrive que certains mots n'appartiennent à aucune grande famille et se trouvent isolés. Ces mots seront qualifiés d'exceptions par opposition aux irrégularités.

On peut donc faire une liste de tous ces lexiques et vérifier que les mots de chaque lexique partagent une même réalité linguistique.

La liste, le contenu et la taille des principaux lexiques linguistiques qui ont été construits à l'ICP à partir du dictionnaire *Le 60 000* de l'ICP sont donnés dans [Belrhali 91]. A quelques rares exceptions, il s'agit de mots d'emprunt ou de mots savants. Par exemple :

- "ch" se prononce [k] dans les mots « savants » (en général d'origine grecque) "trachéal, trachéen, trachéite, trachéostomie, trachéotomie, trachome, trachyte"
- "er" se prononce [e] lorsqu'il est une flexion verbale ("chanter"), ainsi qu'en finale des noms et adjectifs polysyllabiques ("boulanger, léger")
- "er" se prononce [ɛʀ] en finale des noms et adjectifs monosyllabiques, comme dans "cher...", ainsi qu'en finale de mots savants ou d'emprunt, comme dans "africaner..."
- "er" se prononce [œʀ] en finale de mots d'emprunt, comme dans "angledozer, bitter, bookmaker"

Les exceptions sont marginales et causées par des accidents diachroniques ou de faits d'usage. Par exemple :

- "ch" se prononce [ʃ] dans "trachée", son usage plus courant l'a différencié des mots de même origine où "ch"se prononce [k] ("trachéotomie"...)
- "pt" se prononce couramment [pt] dans "dompter", alors que sa prononciation référencée dans le Petit Robert est [t]. Ce mot a subi un accident diachronique : son orthographe s'est transformée arbitrairement de "domter" (du latin "domitare") à "dompter", suite à une analogie au mot "compter"... et l'usage n'a pas suivi la norme de prononciation
  Le Robert oral-écrit fait état de "dilemme" qui se prononce [dilɛm] mais aussi [dilɛmn] comme "indemne" [ɛ̃dɛmn]
- "carrousel", emprunté au XVII^{e} à l'italien "carosello", a conservé la prononciation de s en [s] et se prononce [kaʀusɛl]

Nous n'avons pas inclus dans la grammaire le lexique des mots dont la finale en "-ent" se prononce [ɑ̃], en raison de sa taille trop importante (2 680 mots) par rapport à l'ensemble des autres lexiques. Ce lexique est pour l'instant mis de côté. Il sera remplacé par le lexique des formes verbales fléchies en "-ent", que nous n'avons pas encore obtenu sur *Le 60 000* et qui sera d'une taille bien moindre. La grammaire est actuellement constituée d'une vingtaine de lexiques qui contiennent 1 650 mots et d'un ensemble de 1 270 règles. Le « e » français a un comportement phonétique trés complexe, c'est pourquoi nous lui avons consacré une étude spécifique [Van Eibergen 93]

Bien qu'il soit difficile de décider strictement si une règle définit une exception ou une réécriture régulière on peut estimer que parmi ces règles, environ 600 d'entre elles décrivent à peu près 900 exceptions et les 670 autres caractérisent le mécanisme de base de la réécriture.

Voici quelques exemples de lexiques appelés par les règles :

- Lexique des mots dont l'initiale "ai-" se prononce [e] (et non pas [ɛ]), (taille : 25 mots)
  aider, aigri...
- Lexique des mots dont la finale "-an" se prononce [an] (et non pas [ɑ̃]), (taille : 25 mots)
  barman, carman,...
- Lexique des mots dont la finale "-c" est muette, (taille : 20 mots)
  accroc, ajonc, banc, blanc, broc, caoutchouc...
- Lexique des mots dont la finale "-d" se prononce (mots d'emprunt), (taille : 42 mots)
  apartheid, background, barmaid, baroud...
- Lexique des mots qui contiennent un "-e-" prononcé [e], (taille : 52 mots)
  siderata, referendum, revolver, retriever...
- Lexique des mots dont la finale "-et" se prononce [ɛt], (taille : 16 mots)
  basket, cricket, gadget...cet...
- Lexique des mots dont la finale "-ey" se prononce [ɛ], (taille : 13 mots)
  jockey, bey...

3.2. Le niveau morphophonémique

Les unités lexicales du français sont composées d'affixes et d'une base lexicale au moins : les règles de frontières de mots s'appliquent en général aux frontières de morphèmes. Ainsi un mot graphique devra être considéré, au niveau de la phonétisation, comme autant d'unités morphémiques :

Préfixe + Base

Dans :

- "asocial", "s" précédé du préfixe nominal négatif "a" se transcrit [s]
- "polyacide" "ya" se transcrit [ia] au lieu de [ja] précédé du préfixe nominal poly- signifiant "abondant, nombreux"

Base + Suffixe

On rencontre par exemple le problème de la polyphonie de "tie, tion", où "t", lorsqu'il n'est pas précédé de "s", se transcrit [s] ou [t]. On peut supposer *a priori* que "t" se prononce [s] lorsque "tie" et "tion" sont suffixes.

• La suite "tie" se prononce dans "sotie" et [s] dans "acrobatie"

Mais lorsque ces suffixes sont des flexions, on retombe sur le problème des ambiguïtés catégorielles (que nous verrons plus loin).

Base + Base

L'agglutination est une construction marginale en français, mais on en trouve les traces dans, par exemple :

- "soubresaut" est prononcé comme "soubre" suivie de "saut" : "s" se transcrit [s]
  "tournesol" est prononcé comme "tourne" suivie de "sol" : "s" se transcrit [s]

Ainsi que dans les constructions qui incluent des radicaux grecs fonctionnant comme préfixes, par exemple :

- "immunosuppresseur" est prononcé comme "immuno" suivie de "suppresseur"
- "chimiosynthèse" est prononcé comme "chimio" suivie de "synthèse"

Lorsque l'on analyse des listes tirées du Dictionnaire Inverse de Juilland ou du dictionnaire

Le 60 000 , on s'aperçoit que cette règle ne se vérifie pas toujours. Pour ce type de problèmes, nous proposons des solutions systématiques par lexiques.

- Lexiques des préfixes qui finissent par une voyelle et qui entraînent la transcription de "s" suivant en [s] :
  (a, aéro, anti, auto, bi, carbo, entre, extra, para, re, tétra, tri, ultra, uni, vivi, infra, sulfo, pico, nano, micro, maxi, mini, méga, télé)
  asocial, aérosol, antisocial, autosensoriel, bisulfate, carbosulfure, entresol, extrasensoriel, parasol, resaler, tétrasyllabe, trisyllabe, unisexuel, vivisection, infrason, sulfosel, picoseconde, nanoseconde, microseconde, télésatellite...
- Lexiques des bases préfixales (mots grecs) qui finissent par une voyelle et qui entraînent la transcription de "s" en [s] :
  (éco, immuno, hypo, homo, thio, pyro, turbo, zoo, lacto, phyto, chimio, bio, géo, ortho, psycho)
  écosystème, immunosuppresseur, hyposulfate, homosexuel, thiosulfate, pyrosulfure, turbosoufflante, zoosémiotique, lactosérum, phytosostome, chimiosynthèse, biosynthèse, géosynclinal, orthosystème, psychosocial...
- Lexiques des préfixes qui finissent par "y" et qui entraînent la transcription de "ya" en [ia] :
  (poly, oxy, tachy)
  polyacide, oxyacétylène, tachyarythmie...
- Lexiques des préfixes qui finissent par "i" et qui entraînent la transcription de "ia" en [ia] (et non pas [ija]) :
  (di, bi, tri, anti, vivi, uni, maxi, mini)
  diacétone, biannuel, triannuel, antiarmement, uniaxial...

3.4. Le niveau catégoriel

Du point de vue de la phonétisation, quelques unes des marques graphiques fonctionnelles (dérivations et flexions) sont ambiguës au strict niveau phonotactique. En particulier, les finales "ent", "tions" ont des prononciations différentes selon la catégorie lexicale du mot :

- "ent" ne se prononce pas lorsqu'il est la flexion verbale de la troisième personne du pluriel ("chantent")
- "ent" se prononce [ɑ̃] en finale de noms, d'adjectifs, d'adverbes et de quelques formes verbales ("vent, lent, souvent, sent"...)
- "tions" se prononce [tjɔ̃] lorsqu'il est une flexion verbale ("chantions"...), sauf dans "balbutions, argutions, initions"...
- "tions" se prononce [sjɔ̃] en finale de noms au pluriel ("rations"...)

Si une description linguistique consiste à lever ces ambiguïtés par la valeur catégorielle du mot considéré, une solution pratique est de constituer des lexiques des formes « non verbales », lorsqu'elles ne sont pas homographes. Mais lorsque les deux unités sont homographes, cette ambiguïté catégorielle se double d'une ambiguïté lexicale que ne peut pas lever la grammaire TOPH puisqu'il faut procéder nécessairement à une analyse syntaxique.

Par exemple :

- "er" : "se fier [fje]" (verbe) *vs.* "fier [fjɛʀ]" (adjectif)
- "ent" : "président [pʀezid]" (verbe) *vs.* "président [pʀezidɑ̃]" (nom)
- "tions" : "portions [pɔʀtjɔ̃]" (verbe) *vs.* "portions [pɔʀsjɔ̃]" (nom)
- ambiguïté flexionnelle : "un os [œn-ɔs]" (singulier) *vs.* "des os [dez-o]" (pluriel)

Le lexique des mots dont la finale "-tions" se prononce [sjɔ̃], c'est-à-dire tous les noms (*i.e.* les "non-verbes") contient 1 780 mots (abdications, abductions, aberrations, abjections...). C'est pourquoi, nous avons conservé le lexique contraposé, c'est-à-dire le lexique des verbes.

- Lexique des formes verbales dont la finale est "tions" (prononcée [tjɔ̃], *i.e.* tous les verbes en "-tions" sauf "argutions, balbutions, initions"),
(taille : 825)
  (verbes en "-tir") mentions, partions...
  (verbes en "-ter") abritions, absentions...
- Le lexique des mots dont la finale "-ent" se prononce [ɑ̃] , c'est-à-dire tous les "non-verbes", plus ("sent..."), (taille : 2 680)
  abaissement, abaissement, abattement...

3.5 Le niveau syntaxique

Ces listes sont données à titre informatif, puisque, justement, un lexique de ces formes intégré à TOPH ne peut pas lever les ambiguïtés qui sont d'ordre syntaxique.

- Lexique des homographes dont la finale "-tions" se prononce [tjɔ̃] pour la catégorie verbale et se prononce [sjɔ̃] pour la catégorie des noms :
  (ablations, acceptions, adoptions, affections, attentions, complétions, dations, dictions, éditions, éjections, injections, notions, options, rations, portions, transitions, translations)
- Lexique des homographes où la finale en "-ent" ne se prononce pas pour la catégorie verbale et se prononce [ɑ̃] pour les autres catégories :
  (coïncident, content, couvent, divergent, émergent, expédient, équivalent, évident, excellent, ferment, incident, insolent, négligent, parent, président, récurrent, résident, somnolent, violent)

3.5. Le niveau sémantique

Enfin, il existe un dernier type d'ambiguïté qui ne peut pas être levé par la syntaxe, car il y a homomorphie lexicale et catégorielle. La décision peut être prise alors au niveau sémantique, voire pragmatique. C'est le cas pour le nom :

- "les fils [fis]" (« enfants ») *vs.* " les fils [fil]" (« à coudre »)
- "jet [ʒɛ]" (« jet d'eau ») *vs.* " jet [dʒɛt]"(« avion »)

IV. INTERPRÉTATION DIACHRONIQUE

Certains phénomènes ortho-phoniques sont des héritages de l'histoire de notre écriture. Notre système d'écriture est un système alphabétique qui découle d'un système syllabique (Sémitique Occidental). À son commencement, elle connut deux stades. Le premier

stade est celui où des images expriment une signification : la sémasiographie [Gelb 1973]. Le second, l'écriture restitue directement la langue parlée : la phonographie.

Notre alphabet actuel vient de l'alphabet grec. En effet, les grecs mirent au point un système de voyelles qui articulé aux signes syllabiques en réduisit la valeur à celle de simples signes consonantiques créant ainsi les consonnes.

Il y a trois types d'alphabets en usage, ce qui les distingue c'est la méthode pour indiquer les voyelles.

1 - Grec, latin, etc...
voyelles indiquées par des signes distincts
t - a t -i
2 - Arabe, araméen, hébreu,etc...
voyelles indiquées par des marques diacritiques distinctes
ṭ ṭ
3 - Indien, éthiopien
voyelles indiquées par des marques diacritiques attachées au signe, ou par une modification du signe
t- t.

Notre alphabet actuel connaît donc 6 voyelles, et 20 consonnes graphiques. Chaque lettre de l'alphabet a sa propre histoire.

Exemple : le cas de la lettre "X"

Cette consonne est la seule qui, d'un point de vue phonétique, représente deux sons successifs [ks]. À l'origine, "X" représenté U+S.

X se transcrit de plusieurs façons (référence de la transcription phonétique : Petit Robert 1) :
X se transcrit [k] dans exsangue
X se transcrit [x] dans xérès
X se transcrit [z] dans sixième
X se transcrit [gz] dans existence
X se transcrit [kz] dans xénèlasie
X se transcrit [s] dans soixante
X se transcrit [iks] pour représenter la lettre de l'alphabet
X se transcrit [] dans amoureux

Exemple d'une classe : classe du F

(cer, cle, ner, ser) + f + ("#", s) = [ ]
cerf : 1080 du latin *cervus*
clef : 1080 du latin *clavis*
contreclef : 1754 de contre et de clef
nerf : 1080 du latin *nervus*
serf : XIe du latin *servus*

Ces mots apparaissent en 1080, ce qui correspond à la chanson de geste « La Chanson de Rolland », datée de la fin du XIe. Cette chanson constitue le premier texte de ce genre et de cette importance en langue vernaculaire qui a donné naissance à de nombreux mots.

(neu) + f + (châtel) = [ ]
neufchâtel : 1798 de Neufchâtel

dérivation lexicale, à partir d'un nom propre on a construit un nom commun.

+ ff + = [f]
("#") + f + ("#") = [ɛf]

Cette règle concerne le cardinal de l'alphabet, et utilisée dans les sigles.

+ f + = [f] règle générale

Exemple des différentes règles nécessaires à la phonétisation d'un mot

Voici les différentes règles nécessaires pour la phonétisation du mot "schnauzer".
Classe du s + s + (chnauzer) = [ ]
Classe du c + ch + = [ʃ]
Classe du n + n + = [n]
Classe du a (schn) + au + (zer) = [aw]
Classe du z ("Voyelle") +z+ ("Voyelle") = [z]
Classe du e "Lexique er final [œʀ]") + er + = [œʀ]
Cette règle appelle le lexique suivant :
("Lexique ER final se prononce [œʀ]") = (aft, angledoz, bookmak, boost, bootlegg, bulldoz, ...).

Exemple de phonétisation des mots composés

Le graphème « c » en position finale dans les mots « blanc, croc, franc, porc » ne se prononce pas. Quand ces mots consituent un élément d'un composé le « c » au contact de la voyelle suivante entraîne une liaison. Nous avons donc écrit la règle suivante :
+c+ ("TIRET"+ "VOYELLE") = [k]

Exemple de phonétisation de sigle

Les sigles peuvent être épelés ou lus sans que des règles de prononciation puissent clairement en modéliser le choix [Plénat ??]
« s.n.c.f. » pour la phonétisation de ce sigle, les règles appliquées sont :
("#") +s+ ("#") = (ɛs)
("#") +n+ ("#") = (ɛn)
("#") +c+ ("#") = (se)
("#") +f+ ("#") = (ɛf)

Les emprunts

Un grand nombre de règles d'irrégularités touche un phénomène liguistique bien particulier : les emprunts. Les emprunts contemporains à l'anglais sont très nombreux, l'emprunt aux autres langues est plus occasionnel et est lié à un événement précis :
exemple : pizzeria (italien),

Il existe plusieurs voies d'emprunt, mais la plus importante reste celle des media de communication (presse, télévision, radio). Lorsqu'un mot apparaît oralement par le biais de la télévision (ou de la radio), il est doté d'une prononciation (*e.g.* sniper [snɑjpœʀ]). Parfois, il existe une biphonie pour une même suite de caractère : la suite de caractère « er » (dans les mots anglo-américains, *e.g.* reporter : [ʀəpɔʀtɛʀ, ʀəpɔʀtœʀ]

Un autre exemple d'une double prononciation d'une même graphie, prenons le cas de « u », cette graphie se réalisera d'une manière aléatoire [œ] comme dans hold-up, tantôt [u] comme dans « bulldozer ». Nous avons aussi un cas de polyphonie dans le mot « suspens » qui se réalise de deux manières différentes : [syspɛns, sœspɛns].

Au niveau de la phonétisation, chaque mot doit être considéré selon une évolution qui lui est spécifique.

V. RÉSULTATS ET PERSPECTIVES

Les formes canoniques véhiculent à l'oral des informations des différents niveaux linguistqiues (catégorielles, morphophonémiques) nous avons ajouté dans le lexique une frontière morphologique, et un indice catégoriel (Verbe/Nom ⁀Verbe) par une marque spécifique. Dans ce travail, les aspects touchant à la morphologie dérivationnelle (*e.g.* asocial [asosjal]) n'ont pas été abordés.

Au vu de l'ensemble des irrégularités considérées, nous constatons que la langue est structurée en plusieurs niveaux. La base constituée de l'ensemble des régularités de réécriture (650 règles) est complétée par des lexiques et des règles satellites d'irrégularités (20 lexiques décrivant 1600 mots, et 600 règles décrivant 900 exceptions).

Le dictionnaire *Le 60 000* a été a phonétisé avec cette nouvelle grammaire. Les formes phonétiques sont correctes à 99%. Le taux de phonétisation mal prédit résulte des cas particuliers décrits ci-avant qui ne peuvent pas être résolus par la grammaire tant qu'elle n'intègre pas d'informations de niveau supérieur au lexique.

VI. CONCLUSION

Nous avons voulu mettre en évidence, dans cette étude, le rôle prépondérant que jouent les lexiques. Intégrés dans la grammaire, ils ne doivent pas être considérés comme des simples collections de données mais au contraire comme des représentations naturelles dans la modélisation mixte lexique-règles que propose TOPH.

Le modèle formel TOPH choisi a priori pour modèliser les relations entre chaînes orthographiques et chaînes phonétiques distribue un rôle structurateur équivalent à la description par lexiques et à l'analyse logique par règles. Cette étude s'est attachée à la description la plus approfondie possible des problèmes de la phonétisation du mot. Elle s'est continuée logiquement par l'étude des mots composés [Belrhali *et al.* 1991]. Une seconde étape de ce travail consiste à définir et introduire dans le dictionnaire *Le 60 000* un ensemble de catégories lexicales spécifiques aux phénomènes de l'oral (désambiguïsation de la phonétisation et de l'intonation) [Chatti, 1991]. À partir du lexique des formes canoniques, nous avons généré, en collaboration avec l'équipe TRILAN de Grenoble, le lexique orthographique des formes fléchies (environ 450 000 formes). La généralisation de la grammaire actuelle sur les formes fléchies s'appuiera surtout sur des règles simples et prédictibles de phonétisation des flexions, les bases lexicales irrégulières étant déjà incluses dans les formes canoniques. Le langage TOPH devra donc être étendu afin de prendre en compte le niveau catégoriel parallèlement à la chaîne orthographique.

RÉFÉRENCES

[Aubergé 87] Aubergé V., Contini M., Maret D. & Schnabel B., *TOPH : un outil de phonétisation multilangue*, Bull. de l'Inst. de Phon. de Grenoble, 16, 155-176, 1987.

[Belrhali 91] Belrhali R. & Libert L., *Phonétisation automatique : évaluation et enrichissement d'une grammaire de phonétisation du français*, Mémoire de DEA, Univ. Grenoble-Stendhal, 1991.

[Belrhali 92] Berlrhali R., Libert L., Aubergé V. & Boë L.J., *Organisation de lexiques pour une grammaire de phonétisation du français : "Le 60 000" de l'ICP.*, 19èmes Journées d'Étude sur la Parole, Bruxelles, 1992.

[Catach 84] Catach N., *La phonétisation automatique du français*, Ed. CNRS, Paris, 1984.

[Divay 77] Divay M. & Guyomard M., *Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français*, Thèse de 3ème cycle, Univ. de Rennes, 1977.

[Gelb 73] Gelb I.J. *Pour une théorie de l'écriture.* Flammarion, Idées et recherches. 1973

[Juilland 65] Juilland A., *Dictionnaire phonétique inverse du français.*, Ed. Mouton, The Hague, Paris, 1965.

[Laporte 88] Laporte E., *Méthodes algorithmiques et lexicales de phonétisation de textes, applications au français*, Thèse de 3ème cycle, Univ. Paris VII, 161 1988.

[Le Petit Robert 90] Robert P, *Le Petit Robert 1*, Société du Nouveau Littré, Le Robert, 1969 p, 1990.

[Plénat 93] Marc Plénat, B. Laks, Marc Klein : *L'oralisation des sigles ou : observation sur le mot minimal français.* De natura sonorum Presses universitaires de Vincennes, 1993.

[Prouts 80], Prouts B., *Contribution à la synthèse de la parole à partir du texte, transcription graphème-phonème en temps réel sur microprocesseur*, Thèse de Docteur Ingénieur, Univ. Paris Sud-Orsay, 1980.

[Van Eïbergen 93] Van Eïbergen J., Belrhali R., Étude statistique du « e » bifide. À paraître dans les actes du colloque de Phonologie d'Aix en Provence (Septembre 93).

PREDICTION DE LA VARIABILITE DANS LA DISTRIBUTION DES ACCENTS ET LES DECOUPAGES PROSODIQUES EN FRANCAIS

ELISABETH DELAIS

CNET - LAA / TSS / RCP
22301 LANNION Cedex

ERSS
UA 1033 - CNRS
UNIVERSITE DE TOULOUSE - LE MIRAIL

Résumé

The way the speaker segments a French utterance into prosodic chunks cannot be correctly derived from the linguistic information alone (in particular syntax). Certain phenomena such as rhythmic alternation or syllabic balancing between constituents also play an important role. It seems however difficult to observe and formalise these rhythmic phenomena, that probably result from biological and cognitive constraints, on which we have very little information. This paper presents an approach which attempts to infer from the analysis of a corpus some linguistic constraints and some rhythmic regularities, consequences of these biological and cognitive constraints. The final aim of this work would be to build up a prosodic model based on the interaction and the conflict resolution between linguistic constraints and biological and cognitive constraints.

Mots-clés : Accentuation, Rythme, Prosodie

INTRODUCTION :

La structuration prosodique d'un énoncé est soumise à des contraintes de natures diverses (cf Rossi (1985), Verluyten (1982), Dell (1984) entre autres) :
- linguistiques, dans la mesure où la prosodie remplit des fonctions linguistiques (démarcation des unités syntactico-sémantiques, par exemple)
- cognitives ou biologiques, qui se retrouvent surtout dans la structure rythmique (Fraisse (1974), Pasdeloup (1992). Néanmoins, de nombreux modèles prosodiques (cf. Dell ou Verluyten) s'appuient principalement sur les informations syntaxiques pour générer les structures accentuelles et intonatives des énoncés.

Le travail présenté dans cet article s'intègre dans un projet de recherche dont le but est de construire un modèle prosodique "automatisable" qui puisse générer pour toute phrase la ou les représentations(s) phonologiques correspondantes, en indiquant si possible celle(s) qui serai(en)t préférée(s). Ces représentations phonologiques, qui peuvent aussi être utilisées pour extraire des contours prosodiques à l'aide d'un système d'apprentissage automatique (cf. Tournemire), doivent indiquer :
- quelles syllabes de la suite syllabique doivent être accentuées,
- où se trouvent les fontières des constituants prosodiques : le groupe rythmique (correspondant à peu près au groupe accentuel, Verluyten (1982)) et le constituant intonatif (comparable au groupe intonationnel, Verluyten (1982))

Pour y parvenir, nous avons décidé de surtout travailler sur le rythme. Ce choix est motivé par le fait que le rythme semble relever de principes biologiques et cognitifs essentiels; des études récentes, menées en neuropsychologie (cf. Petitto (1993)), ont en effet montré que les principes rythmiques, qui jouent un rôle important en prosodie, se retrouvent dans d'autres activités humaines tels que les balbutiements des nourrissons ou la mastication... Nous tentons d'observer sur des données expérimentales (corpus enregistré par 4 locuteurs) des régularités de nature rythmique et de nature linguistique (c.à.d. contraintes par la structure morphologique ou syntaxique). Nous essayons ensuite de les formaliser sous forme de principe pour les utiliser dans un modèle prosodique qui s'appuierait sur les interactions et la résolution des conflits entre des contraintes de nature linguistique et de nature biologique et cognitive pour proposer les différentes réalisations possibles d'un même énoncé. Pour construire un tel modèle, nous comptons nous inspirer de"l'optimality theory" (notée OT) de Prince et Smolensky (93), dont les principes sont les suivants :
" [Within OT, the emphasis is on selection of the optimal output forms by a system of ranked, violable constraints]. OT shifts the explanatory burden of linguistic theory from input-based rewrite rules to output-based constraints. Instead of taking an underlying form and transforming it stepwise to

its associated output, in OT it is necessary to allow for the specification of a large a set of *candidate* outputs. The candidate set is evaluated by a system of constraints, which selects the actual output from the available candidates. [...] The central analytical proposal of OT is that constraints are ranked in a hierarchy of relevance. Lower-ranked constraints can be violated in an optimal output form when such violation secures success on higher-ranked constraints." (MacCarthy et Prince (1993), p. 6)

Nous pensons que ce cadre offre des perspectives intéressantes par rapport aux systèmes basés sur des règles de réécriture. Une modélisation élaborée dans ce cadre devrait permettre en effet de proposer plus facilement plusieurs représentations phonologiques possibles pour un même énoncé. Nous émettons d'ailleurs l'hypothèse que les réalisations prosodiques possibles pour un énoncé s'expliquent par la façon dont sont gérés les conflits et les interactions entre des contraintes de nature biologique et cognitive et des contraintes de nature linguistique.

1. METHODE D'ANALYSE :

1.1 CORPUS UTILISE :

Pour tenter de dégager les régularités rythmiques qui peuvent contraindre la structuration prosodique et pour voir comment elles interagissent avec les données linguistiques de nature syntaxique et sémantique, nous travaillons, dans un premier temps, sur des phrases dont la structure syntaxique est simple et le contenu sémantique et pragmatique peu marqué, en vue de neutraliser, autant que possible, dans les productions des locuteurs, les variations qui pourraient résulter de phénomènes d'emphase, de thématisation ou de focalisation.

Le corpus est constitué de quatre phrases porteuses où seul varie le nombre de syllabes (de 3 à 9) d'un syntagme nominal situé en milieu de phrase. Ce syntagme est de la forme [le (adj) (nom)]. Les adjectifs et noms utilisés ont été soigneusement choisis afin d'éviter :

- que le mot ne commence par une syllabe ayant une attaque vide. Ces syllabes ont souvent un comportement particulier (cf. Pasdeloup (1990) et Plénat (1993)).
- que le mot ne contienne un schwa qui puisse être élidé (petit..). Cela pourrait modifier le décompte syllabique.

Les phrases porteuses ont été construites de façon à contraindre et à induire plus ou moins fortement les découpages prosodiques des énoncés. Pour cela, nous nous sommes appuyés sur le poids de la rupture syntaxique et sémantique entre les constituants syntaxiques de surface et sur leur taille. Les phrases sont:

- *Pierre a dit que demain, [le (adj) (nom)], sans doute, irait mieux.*
- *Tu as probablement vu [le (adj) (nom)] que j'ai pris en photo hier.*
- *Le gardien a vu [le (adj) (nom)] de ma voisine.*
- *Les jeunes enfants regardent [le (adj) (nom)] à la télé.*

Le corpus est composé de 261 phrases (87 phrases réitérées trois fois) qui ont été présentées dans un ordre aléatoire pour éviter les effets de liste. Elles ont été lues par 4 locuteurs (2 hommes et 2 femmes). Pour l'instant, nous avons surtout travaillé sur environ 300 phrases, soit un peu moins de la moitié du corpus (2 types de phrases porteuses, 2 réitérations et 3 locuteurs).

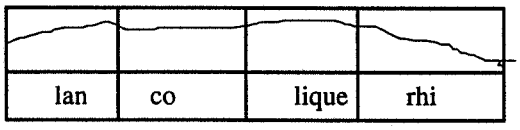
1.2 ANALYSE ACOUSTIQUE DES PARAMETRES PROSODIQUES :

Cette analyse a uniquement pour but de déterminer si les syllabes de la suite syllabique sont accentuées ou non. Pour cette tâche, les indices acoustiques que nous prenons en compte sont la fréquence fondamentale, la durée, et secondairement (en cas d'ambiguïté), l'intensité.

A partir de l'étude des mouvements de la courbe de fréquence fondamentale sur chaque syllabe, nous considérons qu'une syllabe est accentuée dans les cas suivants (pour plus de détails cf. Delais (à paraître)) :

- Montée graduelle de la courbe de F0 sur la syllabe.
- Mouvement descendant de la courbe de F0 et allongement de la durée syllabique (Les accents sont fréquemment réalisés de cette façon en fin de phrase).
- Maxima de F0 et léger allongement de la durée syllabique dans le cas d'une courbe de fréquence fondamentale ayant une forme "hat pattern" (cf. t'Hart et Collier (1975) et Pasdeloup (1990)).

Exemple : *le mélancolique rhinocéros* (accent sur "lique").



Notre étude de la durée syllabique s'appuie sur la durée moyenne des syllabes inaccentuées. Cette durée sert de référence pour juger ensuite si une syllabe est allongée ou non. Pour calculer cette durée, nous nous appuyons principalement sur la durée du noyau vocalique.

A partir de cette analyse, nous construisons pour chaque phrase du corpus une grille accentuelle qui prend la forme suivante :

syllabes	l e	j oe n	z~a	f~a	r e	g a r d	l e	m a	n i	f i k	d r o	m a	d ɛr	a	l a	t e	l e
Contours	→	↗	↘	↗	↘	↗	↘	↗	→	→	↘	→	↗	↘	→	→	↘
Durée (ms)	131	193	173	250	119	227	93	110	140	205	147	123	256	62	94	177	180
Accents	-	+	-	+	-	+	-	+	-	+	-	-	+	-	-	-	+

Phrase N° 91 produite par un locuteur masculin : *Les jeunes enfants regardent le magnifique dromadaire à la télé.*

1.3 INTERPRETATION DES DONNEES :

A partir des informations contenues dans les grilles accentuelles, il est possible de distinguer deux types d'accents en fonction de leur rôle et de leur réalisation :

- Les accents régulateurs rythmiques : Ces accents ont principalement une fonction rythmique, puisqu'ils favorisent la cohésion rythmique de l'énoncé. Leur absence rendrait l'écart entre deux syllabes accentuées trop important et pourrait provoquer la disparition de la perception de rythmicité (cf. Fraisse (1974) et Lieberman (1984)). Les accents régulateurs rythmiques sont caractérisés par un mouvement mélodique montant de faible ampleur, et parfois par un allongement non significatif de la durée syllabique. Ils se situent généralement sur la syllabe initiale ou sur l'antépénultième d'un mot lexical (cf accent secondaire, Fonagy (1979) et Pasdeloup (1990), ictus mélodique, Rossi (1985)). Dans la grille que nous avons présentée, les accents sur "j oe n" et sur "ma" sont régulateurs rythmiques.

- Les accents démarcatifs de constituants prosodiques : Ces accents ont une fonction linguistique. Ils marquent en effet la frontière droite des constituants prosodiques, qui se superposent généralement, mais pas nécessairement, à des constituants syntaxiques. Les accents démarcatifs se caractérisent par un mouvement montant ou descendant assez ample de la courbe de fréquence fondamentale et par un allongement significatif de la durée syllabique. Ils se situent toujours sur la dernière syllabe d'un mot lexical (sauf si celle-ci a pour noyau vocalique un schwa) et correspondent généralement à un contour prosodique de fin de phrase ou de syntagme (cf. accent primaire, Pasdeloup (1990), accent final, Mertens (1987 et 1992) Di Cristo (sous presse), accent interne, Rossi (1985). Dans la grille présentée au § 1.2, les accents qui portent sur "f ~a", "gard", "fik", "dɛr" et "le" sont démarcatifs de constituants prosodiques.

En fonction de l'ampleur des mouvements mélodiques et des allongements syllabiques des accents démarcatifs, il est possible de définir des constituants prosodiques de deux niveaux hiérarchisés :
- Pour la phrase présentée, le découpage en constituants de niveau 1, appelés groupes rythmiques (cf. groupe accentuel, Mertens (1992) et Verluyten (1982), mot rythmique, Pasdeloup (1990), unité rythmique, Hirst et Di Cristo (1993)) est :
(Les jeunes enfants) (regardent) (le magnifique) (dromadaire) (à la télé).
- Les accents démarcatifs portant sur les syllabes "f ~a", "dɛr" et "le" marquent, en revanche, la frontière de constituants prosodiques de niveau supérieur que nous appelons constituants intonatifs (cf. séquence rythmique, Pasdeloup (1990), groupe intonatif, Mertens (1993), unité intonative, Di Cristo (1993 et sous presse), groupe intonationnel, Verluyten (1982)). Nous obtenons donc le découpage en constituants intonatifs suivant :
- [(les jeunes enfants)] [(regardent) (le magnifique) (dromadaire)] [(à la télé)].

Nous avons opté pour une définition fonctionnelle de l'accent plutôt que pour une définition formelle basée sur sa place dans le mot (cf. Mertens (1992), Pasdeloup (1990)). Ce choix est motivé par le fait que certains accents, portant sur la dernière syllabe d'un mot lexical ou d'un constituant syntaxique de surface, ont davantage un rôle de régulateurs rythmiques. Ils existent cependant des cas limites où il est difficile de savoir si de tels accents sont régulateurs rythmiques ou démaractifs de constituants prosodiques.

Les résultats obtenus pour la distribution accentuelle et le découpage prosodique sont ensuite vérifiés à l'aide d'une écoute attentive de chaque phrase et par un procédé de "mimicking".

2. RESULTATS OBTENUS :

La distribution des accents et les découpages prosodiques obtenus pour chaque énoncé ne sont pas toujours identiques dans les productions des trois locuteurs que nous avons étudiées. Malgré ces différences, certains phénomènes apparaîssent très fréquemment. Ils peuvent donc servir de point d'appui pour dégager les régularités rythmiques et pour formaliser des principes ou contraintes de nature biologique et cognitive. En outre, la distribution des accents et la formation des groupes rythmiques et des constituants intonatifs nécessitent la prise en compte d'informations lexicales et syntaxiques. Il est donc aussi possible de définir des contraintes linguistiques.

2.1 REGULARITES RYTHMIQUES :

Les régularités de nature rythmique que nous avons relevées concernent la fréquence des accents, les écarts entre accents démarcatifs et la recherche d'un équilibre syllabique entre les constituants prosodiques (groupes rythmiques et constituants intonatifs).

- La fréquence des accents :

Sur l'ensemble des données étudiées, nous avons calculé la part des syllabes accentuées, en distinguant celles portant un accent démarcatif et celles portant un accent régulateur. Pour chacun des locuteurs, nous obtenons des résultats assez semblables :

	syllabes inaccentuées	accents régulateurs	accents démarcatifs
Locuteur 1	59 %	12 %	29 %
Locuteur 2	61 %	11 %	28 %
Locuteur 3	62 %	11 %	27 %

Bien que les trois locuteurs ne réalisent pas toujours leurs accents à la même place, l'écart moyen entre accents (démarcatifs ou régulateurs) est quasi identique. Il se situe entre 1,42 et 1,54 syllabes, c'est à dire qu'il y a un accent toutes les deux ou trois syllabes (cf. Pasdeloup (1990) et le "least syllable principle" de Wenk et Wioland (1982)). Dans plus de 80 % des cas, cet écart est respecté. Il semble donc important d'en tenir compte pour formaliser les principes ou contraintes rythmiques. En outre, il est nécessaire d'étudier en détails les 20 % des cas où cet écart n'est pas respecté afin d'en comprendre la raison (par exemple, respect de certaines contraintes linguistiques).

Sur l'ensemble de nos données, nous n'avons jamais rencontré deux syllabes contiguës accentuées (cf Dell (1984), Pasdeloup (1990), et pour une position contraire Mertens (1992)). Nous pouvons donc poser la contrainte suivante :

CONT : deux syllabes contiguës ne peuvent pas être accentuées.

- L'écart entre accent démarcatif ou la taille des groupes rythmiques :

Les groupes rythmiques, dont la frontière est indiquée par un accent démarcatif, contiennent en moyenne 3 ou 4 syllabes (3,5 syllabes) chez chaque locuteur. Sur l'ensemble des énoncés étudiés :

- le groupe rythmique le plus petit a une syllabe :

Pierre dans (Pierre) (a dit que demain) (le beau chien) (sans doute) (irait mieux).

- les plus grands ont six ou sept syllabes :

Tu as probablement vu dans (Tu as probablement vu) (le beau singe) (que j'ai pris) (en photo) (hier).

Lorsque les groupes rythmiques contiennent plus de trois syllabes, un accent régulateur rythmique y est très souvent réalisé (cf. Wenk et Wioland (1982)). Ainsi, dans la phrase *"Les jeunes enfants regardent le magnifique singe à la télé"*, dont le découpage en groupes rythmiques est le suivant :

- (les jeunes enfants) (regardent) (le magnifique singe) (à la télé),

le groupe "(le magnifique singe)" est toujours réalisé dans notre corpus d'une des deux façons suivantes :

le	ma	gni	fique	singe
-	-	+	-	+
-	+	-	-	+

D'après nos observations, l'écart entre les accents démarcatifs visent à ce qu'un groupe rythmique ne contiennent pas plus de six (exceptionnellement sept) syllabes (cf Pasdeloup (1992)). Ainsi, lorsqu'un groupe rythmique contient plus de 6 syllabes, il est presque toujours décomposé en deux groupes distincts. Lorsque plusieurs types de découpages sont possibles, on retiendra, autant que possible, ceux qui respectent l'integrité des constituants syntaxiques de surface. Par exemple, pour la phrase *"le gardien a vu le gentil chamois de ma voisine"*, le découpage qui suit n'a jamais été observé, bien qu'il respecte la contrainte de taille :

* (le gardien) (a vu le gentil) (chamois) (de ma voisine).

Nous pouvons donc formuler une contrainte pour rendre compte de la taille maximale du groupe rythmique :

MAX : Un groupe rythmique ne peut pas avoir plus de 7 syllabes.

- La recherche d'équilibrage syllabique :

D'après nos observations, la recherche d'un équilibrage syllabique entre les constituants prosodiques exerce des contraintes fortes. Les découpages prosodiques obtenus pour les phrases analysées indiquent en effet que le locuteur découpe son énoncé et réalise ses accents de façon à ce que le nombre de syllabes de chaque constituant intonatif soit à peu près équivalent dans une phrase. Si nous comparons les découpages obtenus pour deux phrases ayant la même structure syntaxique, mais dont le nombre de syllabes d'un des constituants varie, nous obtenons des résultats différents.

Soient les phrases :

- *Le gardien a vu le beau chien de ma voisine.*
- *Le dernier fils du gardien a vu le beau chien de ma voisine.*

Les découpages prosodiques obtenus sont respectivement :

- [(le gardien)] [(a vu) (le beau chien)] [(de ma voisine)]
- [(le dernier fils) (du gardien)] [(a vu) (le beau chien) (de ma voisine)]

Pour chacun des énoncés, le nombre de syllabes de chaque constituants intonatifs est à peu près équivalent. Pour le premier, leur taille est de trois, cinq et quatre syllabes. Pour le second, elle est de sept et de neuf syllabes. La recherche d'un équilibrage syllabique entre les constituants prosodiques semblent donc intervenir dans les choix des découpages, même si cela implique un non respect de la hiérarchie syntaxique (cf. Martin (1987)). Ainsi, dans la première phrase, bien que les constituants "*le beau chien*" et "*de ma voisine*" soient très liés dans la hiérarchie syntaxique, ils sont fortement disjoints prosodiquement. Dans quatre des six réalisations de cet énoncé observées, il y a même une pause entre *chien* et *de ma voisine*.

2.2 CONTRAINTES LINGUISTIQUES :

Une étude approfondie des réalisations obtenues montrent que des principes linguistiques, qui reposent principalement sur la catégorie grammaticale des mots et sur la structure syntaxique de la phrase, contraignent le placement des accents et les découpages prosodiques possibles pour un énoncé donné. Ces principes ne sont pas nécessairement respectés, surtout s'ils entrent en conflit avec ceux mentionnés au § 2.1. Cependant, ils peuvent très souvent permettre d'expliquer les réalisations qui ne répondent pas aux régularités rythmiques. Les contraintes linguistiques interviennent sur la distribution des accents et sur les stratégies de découpage en constituants prosodiques (constituants intonatifs et groupes rythmiques).

- La distribution des accents :

Sur l'ensemble des données étudiées, il n'y a aucun accent sur un mot grammatical. Nous pouvons donc poser le principe suivant (cf. aussi Pasdeloup (1992), Mertens (1993)), qui s'applique dans un style neutre en lecture :

GRAM : les mots grammaticaux ne peuvent pas porter d'accent.

L'étude d'un corpus spécifique doit être envisagée afin de savoir quels mots entrent réellement dans la catégorie des mots grammaticaux. Certaines prépositions (*"pendant"*, par exemple) ou auxiliaires (les formes polysyllabiques) peuvent sans doute être accentuées (cf. Mertens (1993) et le trait [± clitique]).

Cette contrainte peut expliquer pourquoi il existe des suites de trois ou quatre syllabes consécutives inaccentuées, c'est à dire qui ne respectent pas le principe rythmique de fréquence entre accents. Ainsi, dans une des phrases porteuses que nous avons étudiées (" *Le gardien a vu [le Adj Nom] de ma voisine.*", il y a toujours trois syllabes consécutives inaccentuées dans le constituant (de ma voisine). Le seul schème accentuel possible dans ces cas est en effet:

de	ma	voi	sin(e)
-	-	-	+

- Les stratégies de découpages prosodiques :

Un étude approfondie des découpages obtenus et de la localisation des accents démarcatifs laisse supposer que certaines contraintes (qui visent à préserver l'integrité des constituants syntaxiques de surface) jouent un rôle dans la formation des constituants prosodiques. Ces contraintes, d'après les observations, se conçoivent principalement en terme d'alignement entre constituants prosodiques et syntaxiques.

Nous voulons donc essayer de les formaliser, en s'appuyant sur le principe de "Generalized Alignment" (Mac Carthy et Prince (1993) p. 2) qui a été formulé dans le cadre de "l'optimality theory" :

GA : Align (Cat1, Edge1, Cat2, Edge2) = def
∀ Cat1 ∃ Cat2 such that Edge1 of Cat1
and Edge2 of Cat2 coincide
Where : Cat1, Cat2 ∈ PCat ∪ GCat
Edge1, Edge2 ∈ {Right, Left}

Sur l'ensemble du corpus, les groupes rythmiques s'alignent presque systématiquement sur les constituants syntaxiques de surface (sauf s'ils contiennent plus de 6 ou 7 syllabes). Pour construire le constituant syntaxique de surface, nous proposons d'utiliser la règle de construction de la phrase phonologique (Nespor et Vogel (1986)) :
"i) Adjoindre au Nom, Verbe ou Adj de tête tous les items qui le précédent à l'intérieur du syntagme.
ii) Adjoindre toute préposition et / ou tout item non lexical qui reste (conjonction, complémenteur ...) au premier N, V ou Adj de tête qui suit."

Nous appelons le groupe ainsi formé groupe de sens (GS). A partir de ce constituant construit sur des bases syntaxiques, il est possible de poser les principes d'alignement entre constituants prosodiques et groupes de sens suivants (présentés selon une hiérarchie de pertinence) :
INTONATIF : Align (CI, r, GS, r)
RYTHMIQUE : Align (GR, r, Wordlex, r)
RYTHM : Align (GR, r, GS, r), classé après MAX, § 2.1.

Ces principes sont respectés, à moins d'un conflit avec les principes rythmiques. Une étude spécifique devra donc être menée afin de les hiérarchiser entre eux et en fonction des principes rythmiques.

La rupture syntaxique qui semble fréquemment respectée est la séparation entre le syntagme nominal

sujet et le groupe verbal. Elle semble donc exercer une contrainte sur la formation des constituants intonatifs, qui peut aller à l'encontre d'une recherche d'équilibre syllabique entre les constituants (cf § 2.1). De ce fait, nous obtenons des réalisations où l'équilibrage syllabique entre les constituants prosodiques n'est pas satisfait :

- [(les jeunes enfants)] [(regardent) (le magnifique chien)] [(de ma voisine)]

Sur l'ensemble des énoncés étudiés, il existe nénamoins quelques cas (moins de 10 %) où la rupture entre le syntagme nominal et le syntagme verbal n'est pas respectée. La recherche d'équilibre syllabique semble donc motiver les découpages prosodiques obtenus. Regardons, par exemple :

- [(les jeunes enfants) (regardent)] [(le gros rhinocéros)] [(à la télé)].

Nous n'avons malheureusement pas assez de données allant de ce sens pour pouvoir formuler des hypothèses précises sur les mécanismes mis en oeuvre pour résoudre les conflits qui résultent de la recherche d'un équilibre syllabique et du respect de cette rupture.

CONCLUSION ET PERSPECTIVES :

Les résultats obtenus permettent de dégager des régularités de nature rythmique (fréquence des accents, taille des constituants prosodiques et recherche d'un équilibrage syllabique) et des contraintes linguistiques (alignement des constituants prosodiques et syntaxiques, rupture entre le syntagme nominal sujet et le SV). Ces phénomènes semblent contraindre les réalisations prosodiques possibles.

Afin de mieux définir le poids de ces contraintes linguistiques et rythmiques, et de comprendre comment elles intéragissent, nous comptons poursuivre le même type d'analyse sur l'ensemble du corpus, ainsi que sur les productions d'autres locuteurs et sur des énoncés moins contraints linguistiquement. Cela devrait permettre de mieux formaliser ces contraintes de nature diverses et de les hiérarchiser afin de rendre compte des mécanismes mis en oeuvre pour résoudre les contradictions entre les différents types de contraintes et de ne retenir que les représentations bien formées.

Un modèle prosodique élaboré dans ce cadre précis aurait un intérêt certain pour la synthèse de la parole, dans la mesure où il pourrait permettre de générer plusieurs représentations phonologiques possibles pour un énoncé (l'évaluation des représentations générées se ferait par rapport au poids des contraintes respectées ou non). Cela devrait donc introduire de"la variabilité prosodique" (cf. Tournemire (à paraître)).

Bibliographie :

DELAIS, E. (à paraître) "Structure rythmique et prosodique du français : pour une approche psychocognitive", *LINX , numéro 29 (février 94)*, Paris X - Nanterre.

DELL, F., 1984, "L'accentuation dans les phrases en français", *Forme sonore du langage : structure des représentation en phonologie*, F.Dell, D. Hirst et J.R Vergnaud (eds); Hermann, Paris, pp 65-122.

DI CRISTO, A. (à paraître) "French Intonation", *Intonations systems*, D.Hirst et A.Di Cristo (eds), Cambridge University Press, Cambridge

DI CRISTO, A et D. HIRST ; 1993, "Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du Français", *Travaux de l'Institut Phonétique d'Aix en Provence N°15.*

FONAGY, I.; 1979, L'accent français : accent probabilitaire, l'accent du français contemporain in *Studia Phonetica N° 15*, pp 123-233

FRAISSE, P. , 1974, *Psychologie du rythme*, Presses Universitaires de France, Paris, 360 p.

LIEBERMAN, Ph. , 1984, *The Biology and Evolution of Language*, Harvard University Press, Cambridge

MacCARTHY, J. et PRINCE, A., 1993, "Generalized Alignment", manuscrit, University of Massachusetts, Amherst et Rutgers University.

MARTIN, P. ; 1987, Structure rythmique de la phrase française, statut théorique et données expérimentales, *Actes des 16èmes Journées d'Etude sur la Parole*, Hammamet, 5-9 Octobre 1987, 225-258

MERTENS, P.; 1987, *L'intonation du français. De la description linguistique à la reconnaissance automatique*, Thèse de Doctorat, Université de Leuven, p 917

MERTENS, P.; 1992, "L'accentuation de syllabes contiguës", *ITL Review of Applied Linguistics*, 95-96, p145-165

NESPOR, M. et VOGEL, I. , 1986, *Prosodic Phonology*, Foris, Dordrecht.

PASDELOUP, V. , 1990, *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*, Thèse de Doctorat, Université d'Aix en Provence, Aix-Marseille

PASDELOUP, V. , 1992, "A prosodic Model for French Text -to - speech synthesis : A psycholinguistic approach", *Talking Machines : Theories, Models, and Designs*, Bailly G, C. Benoit et T.R Sawallis (eds), Elsevier Science Publishers, pp 335-48

PETITTO, L.A. , 1993, "On the ontogenetic requirements for early language acquisition", *Developmental neurocognition : Speech and face processing in the First year of life*, B. Boysson-Bardies, S. de Schonen, P. Jusczyk, P. Mcneilage et J. Morton (eds), Kluwer Academic Publishers, pp 365-84

PLENAT, M. (à paraître) " L' "extramétricité " des voyelles initiales", *Poceedings de la Conference of the association for French Language Studies*, 16-17 septembre 1993, Université d'Aix en Provence.

PRINCE, A. et SMOLENSKY, P., 1993, "Optimality Theory : constraints interaction in Generative grammar", Manuscrit, Rutgers University, Newbrunswick et University of Colorado, Boulder.

ROSSI, M.; 1985, "L'intonation et l'organisation de l'énoncé", *Phonetica N°42*, pp 135-153

'T HART, J. et R. COLLIER , 1979, "On the interaction of accentuation and intonation in Dutch", *Proceedings of the IXth Int. Congr. Phon. Sci.*, 2, Copenhague, pp 395-402

TOURNEMIRE de, S. (à paraître) "Recherche d'une stylisation extrême des contours de F0 en vue de leur apprentissage automatique", dans ces Actes.

SELKIRK, E. , 1986, "On derived Domains in Sentence Phonology", *Phonology Yearbook 3*, pp 371-405

VERLUYTEN, S.P , 1982, *Investigation on French Prosodics and Metrics*, Phd Dissertation, Anvers, Belgique

WENK, B. et WIOLAND, F. ; 1982, Is French really syllable-timed?, *Journal of Phonetics 10*, Academic Press, Londres, pp 193-216

INTERET DU TEST DE RIMES DANS L'EVALUATION DE L'INTELLIGIBILITÉ EN PRÉSENCE DE BRUIT MASQUANT

J.M. BENE*, C. DUBREUIL*, J.C. DUCLOS*
J.C. NORMAND°, M VERNET #, G. PACHIAUDI #

* Service d'ORL, Centre Hospitalier Lyon-Sud, 69310, Pierre Bénite, France
° Service de Médecine du Travail, Centre Hospitalier Lyon-Sud, 69310, Pierre Bénite
Institut National de Recherche sur les Transports et leur Sécurité, Av S. Allende, 69500, Bron, France.

Résumé

La mise au point de tests d'intelligibilité est l'étape préalable au développement de la communication verbale à l'intérieur des véhicules, notamment dans l'optique d'une amélioration de l'interface homme-machine. En matière de transports, le canal visuel est de plus en plus saturé de par le grand nombre d'informations qu'il véhicule. Il serait souhaitable de faire passer des informations par le canal auditif. Mais les véhicules sont soumis à un important bruit de fond. Une méthode permettant d'apprécier l'intelligibilité dans le bruit de fond des véhicules, de façon simple et reproductible, serait donc bienvenue.

Les auteurs ont étudié la place du Test de Rimes dans l'évaluation de l'intelligibilité sur une population malentendante (surdités de perception, surdités de transmission, cophoses unilatérales).

Le Test de Rimes, ou Test par paires minimales (VOIERS, 1967) a été adapté pour le français par PECKELS et ROSSI (1973) puis mis en oeuvre par CARTIER et ROSSI (1975). Il est fondé sur l'pposition de traits acoustiques au sein de paires minimales, selon la matrice de JAKOBSON *(JAKOBSON, FANT, HALLE, 1963, (4))*.

Dans une première partie, les résultats du Test de Rimes sont comparés aux résultats des tests paracliniques habituellement utilisés dans l'exploration des surdités (audiométries tonale et vocale).

Dans une seconde partie, la dégradation de l'intelligibilité est étudiée par le Test de Rimes en présence de deux bruits masquants, correspondant à deux véhicules différents, roulant à des vitesses différentes sur deux revêtements différents. La dégradation de l'intelligibilité est étudiée de façon globale puis de façon qualitative au niveau de chaque trait acoustique.

Mots-clés : Surdité, Bruit masquant, Audiométrie, Test de Rimes.

PRESENTATION DE L'ETUDE

Cette étude s'est attachée à étudier les moyens d'apprécier l'intelligibilité dans les véhicules en marche sur une population malentendante. En effet, l'environnement acoustique est un facteur de confort et de sécurité à l'intérieur des véhicules (voitures, mais aussi trains, avions etc...) *(GILLOIRE, LOCKWOOD, BOUDY, 1990, (3))*.

De nombreuses actions en matière de réduction du bruit ont contribué à abaisser les niveaux sonores provenant des différentes sources. Il n'empêche que les niveaux résiduels dans l'habitacle demeurent une gêne pour la communication entre passagers, et un frein au développement des systèmes vocaux dans les véhicules. Comme préalable à l'amélioration de l'intelligibilité dans le bruit de fond des voitures, il convenait de mettre au point une méthodologie d'appréciation de l'intelligibilité. De nombreuses méthodes d'appréciation de l'intelligibilité existent. Elles sont utilisées en pratique courante en médecine, en recherche ou dans l'industrie. Parmi ces méthodes, on s'est attaché à montrer l'interrêt tout particulier du *Diagnostic Rhyme Test* de VOIERS.

Parmi les avantages que présente ce test, la possibilité de l'enregistrer sur bande magnétique ou autre support sonore, et d'en dépouiller les résultats par informatique le rendait très attractif de par sa facilité d'utilisation.

Ce test avait déja été étudié auparavant sur une population aux capacités auditives normales *(PACHIAUDI, VERNET, FOTI, 1987, (6))*. Dans cette étude, on s'est attaché à en montrer l'intérêt dans une population malentendante. En effet, les sujets malentendants sont de plus en plus nombreux, ne serait-ce que du fait des presbyacousies dont le nombre augmente avec l'espérance de vie.

Cette étude a été effectuée en deux temps :

1) Une première partie a été effectuée en atmosphère silencieuse. Sur une population malentendante a été realisé le Test de Rimes, ainsi qu'une dictée de 50 chiffres, et des audiométries

tonales et vocales. Ceci a permis de voir que ce mode d'appréciation de l'intelligibilité est utilisable dans une population malentendante, et a permis de connaître les informations apportées par ces tests par rapport aux informations données par les explorations fonctionnelles paracliniques utilisées couramment dans l'exploration des surdités.

2) Une deuxième partie a été effectuée en présence de différents bruits masquants correspondant à différentes situations de vitesse, de revêtements de routes et de véhicules. Chaque situation a été reproduite en laboratoire après enregistrement "in situ" sur bande magnétique. Le but est alors d'étudier la dégradation de l'intelligibilité telle que la caractérise le Test de Rimes, et, par l'étude de l'évolution dans le bruit des différents traits acoustiques, de chercher à alléger le test.

MATERIEL ET METHODE

Trois types de surdité ont été étudiés : surdités de perception, surdités de transmission, cophoses unilatérales.

Dans une première partie, en **atmosphère silencieuse**, sur chaque sujet ont été réalisés successivement une audiométrie tonale, une audiométrie vocale avec les listes de LAFON (4) (effectuées de façon binaurale et sans bruit masquant), un Test de Rimes (Test de VOIERS traduit par ROSSI, complété d'un test par triplets), une dictée de chiffres. Le Test de Rimes a été enregistré sur bande magnétique par trois locuteurs à une intensité moyenne de 61 dB.

On a retenu 35 témoins, se répartissant en 16 sujets atteints de surdité de perception, 9 sujets atteints de cophose unilatérale, 10 sujets atteints de surdité de transmission.

Dans une deuxième partie, deux types de **bruits masquants** ont été sélectionnés. Ces deux bruits correspondent à deux véhicules différents roulant à des vitesses différentes sur des routes différentes. Les bruits ont été enregistrés de façon stéréophonique au niveau de chacune des oreilles du conducteur. L'intensité du bruit masquant est variable selon les fréquences, s'échelonnant de 90 à 40 dB des graves vers les aigus. Les deux situations ont un écart de 10 dB en moyenne.

Chacune des surdités a été explorée de la même façon que dans la première partie. Puis, dans chaque situation bruitée et en silence, on a effectué un test par doublets, un test par triplets et une dictée de chiffres.

On a exploré 41 témoins, se répartissant en 23 sujets atteints de surdité de perception, 9 sujets atteints de cophose unilatérale, 9 sujets atteints de surdité de transmission.

L'analyse statistique des données a été effectuée avec des logiciels SPSS / PC+ et Epi-Info (CDC Atlanta - OMS Genève). Elle a fait appel, pour les données continues, à l'analyse de variance en cas de distribution normale des variables, et aux tests de Kruskal-Wallis et de Friedman dans le cas contraire. Le risque de significativité retenu pour les analyses (risque de première espèce) est de 5 %. Les résultats sont exprimés en moyenne plus ou moins écart-type.

PARAMETRES ETUDIES

L'audiométrie tonale a consisté en la mesure du seuil à 250, 500, 1000, 2000, 4000 et 8000 Hz. On en a tiré un *indicateur de perte sur les graves* (moyenne de la perte sur les 3 fréquences graves) et un *indicateur de perte sur les aigus* (moyenne de la perte sur les 3 fréquences aiguës)

La **capacité auditive théorique** se définit comme la probabilité de la discrimination phonétique, en fonction de l'intensité, octave par octave *(DUCLOS, 1986, (2))*. Elle se calcule à partir de l'audiogramme tonal d'une part, et d'un abaque dont les valeurs ont été déterminées expérimentalement d'autre part. Elle permet une estimation de la probabilité d'erreurs phonétiques.

A partir de l'**audiométrie vocale**, nous avons retenu comme unique paramètre le score d'erreurs à la liste dite à 70 dB; ceci au risque de restreindre de façon drastique les informations notamment qualitatives qu'est susceptible d'apporter le Test Phonétique.

Pour le **test de VOIERS**, on prend en compte :
le *score global de VOIERS*, pourcentage de réponses justes aux 468 stimuli correspondant à autant de paires minimales ;
les *scores partiels par trait*, correspondant au pourcentage de réponses justes aux stimuli relatifs à chacun des cinq traits testés.

Pour les triplets, les résultats sont donnés par:
le *nombre total d'erreurs* aux 93 stimuli, et
le *détail des erreurs* d'un trait vers l'autre.

Dans les deux cas, on prend en compte le *nombre de non-réponses*.

L'**âge** a été pris en compte.

Pour la **dictée de chiffres**, le pourcentage de réponses fausses a été retenu (*score de la dictée de chiffres*).

RESULTATS

En **atmosphère silencieuse**, les résultats sont très différents selon les types de surdité.

Pour les sujets atteints de surdité de perception, il existe une corrélation linéaire entre le Test de Rimes et la perte audiométrique sur les aigus (R = 0,63), la CAT (R = 0,74) et la vocale (R = 0,66).

Pour les sujets atteints de cophose unilatérale, le Test de Voiers est corrélé avec la perte sur les aigus. Par ailleurs, plus le sujet est âgé, plus il fait d'erreurs à la dictée de chiffres.

Pour les sujets atteints de surdité de transmission, les résultats sont différents des autres types de surdité : le Test de Rimes et la dictée de chiffres suivent des évolutions parallèles.

En **atmosphère bruitée**, les résultats de la dictée de chiffres n'ont pas été exploitables : ils se sont limités à quelques chiffres isolés ou à une feuille blanche dans la plupart des cas.

Les non-réponses se répartissent de façon homogène entre les trois types de surdité : il n'existe pas de différence significative.

Pour ce qui est des résultats globaux au test de VOIERS, il existe une différence de comportement entre les sujets atteints de différents types de surdité: pour les sujets atteints de surdité de perception et de cophose unilatérale, lorsque le bruit augmente, le score global de VOIERS et le nombre d'erreurs aux triplets augmentent de façon statistiquement significative, l'effet étant le plus net pour les sujets atteints de surdité de perception. Par contre, pour les sujets atteints de surdité de transmission, on ne peut pas le mettre en évidence, il n'existe pas de différence significative entre les trois situations de bruit.

DISCUSSION

1) Etude en atmosphère silencieuse

Il est possible de comparer les informations données par le Test de Rimes de VOIERS et le Test Phonétique de Lafon. Même si ces deux tests permettent de mesurer l'intelligibilité, ils n'ont pas été élaborés dans le même but : l'un (VOIERS) est destiné à tester des appareils de codage, de synthèse ou de transmission de la parole; l'autre est destiné à apprécier la qualité d'un sujet à identifier une structure verbale. Dans les deux cas, on s'est attaché à étudier l'intelligibilité au niveau segmental, en ne prenant pas en compte toutes les informations supra-segmentales.

Le mot de trois phonèmes, utilisé dans les deux tests, est un compromis entre la division ultime de la langue, le phonème, et des ensembles linguistiques plus importants : mots plus longs ou phrases. On peut ainsi bénéficier d'un lexique abondant dans la langue, avec des mots suffisamment courts pour qu'un défaut d'intelligibilité ne soit pas complètement compensé par les redondances de la langue, la compensation étant d'autant plus facile que l'ensemble linguistique est plus long.

Dans le Test de Rimes, on ne s'est pas attaché à reproduire, au niveau de la fréquence des phonèmes testés, la fréquence de ces phonèmes dans la langue, contrairement à ce qui est pratiqué dans le Test Phonétique, puisque ce n'est pas l'équilibre phonémique qui est recherché, mais l'examen de toutes les oppositions de traits de la langue.

Dans les deux cas, la prosodie n'apporte pas d'information, du fait de la petite taille du matériel phonétique utilisé. Les sujets ne peuvent pas être aidés non plus par le sens d'une phrase qui dans les deux cas n'existe pas. On supprime donc l'influence de ces deux niveaux de redondance suprasegmentale qui sont pourtant utilisés par les malentendants pour compenser les défauts d'intelligibilité au niveau segmental.

2) Etude en atmosphère bruitée

Le nombre de non-réponses a augmenté avec les bruits masquants plus élevés. L'augmentation n'est significative que dans le cas des perceptions. Le nombre de non-réponses ne peut donc pas être considéré comme un indicateur d'intelligibilité. On peut distinguer deux types de comportement parmi les sujets : le sujet "risque-tout" qui dans le doute met toujours une réponse au risque qu'elle soit fausse, et le sujet plus timoré qui dans le doute préfère s'abstenir. Certains ont même dit, en fin de test, avoir coché une partie au hasard afin de ne pas perdre le fil du test.

Le rapport signal/bruit diminue entre la situation non bruitée et les deux situations bruitées. On trouve une diminution nette du score globale de VOIERS dans le groupe des perceptions, et une diminution légèrement moindre, mais toujours significative dans le groupe des cophoses unilatérales. Par contre, on ne retrouve pas de différence significative dans le groupe des transmissions. La faiblesse des effectifs en est une explication plausible.

Par l'étude des scores par trait et par le test par triplets, on a une vision analytique des traits atteints en priorité ou atteints plus tardivement. On n'a pas pu mettre en évidence de différence d'atteinte entre les différents traits au niveau statistique. Là encore, cette étude mériterait d'être complétée par une étude portant sur des populations plus importantes et plus homogènes.

CONCLUSION

Si la possibilité d'utiliser les Test de Rimes comme moyen d'appréhender l'intelligibilité en présence de bruit masquant chez des malentendants a été montrée, il persiste quelques points que cette étude n'a pas totalement élucidés.

Une étude plus qualitative permettrait de montrer quels types d'informations apportent respectivement des tests comme celui de LAFON et le Test de Rimes.

L'étude de plusieurs types de bruits masquants permettrait peut-être de montrer quels traits de la matrice phonologique de JAKOBSON sont atteints selon les spectres des bruits.

De même, des effectifs plus importants permettraient de savoir quels sont les traits dont l'atteinte est la plus significative en matière d'intelligibilité. On pourrait alors envisager l'utilisation de quelques traits pertinents seulement, de façon à avoir un test plus rapide.

BIBLIOGRAPHIE

(1) CARTIER, M., ROSSI, M., 1975, Le test de diagnostic par paires minimales : mise en oeuvre et résultats, Travaux de l'Institut de Phonétique d'Aix, volume 2, 1-30.

(2) DUCLOS, J.C., 1986, Du bruit, de la surdité et de l'information (aspects physiologiques et psychophysiologiques), Thèse d'état de doctorat ès sciences, Université LYON 1.

(3) GILLOIRE, A., LOCKWOOD, P., BOUDY, J., 1990, Problèmes liés à la communication dans les véhicules, Journées spécialisées INRETS "communication dans les véhicules" du 31 01 1990. Actes INRETS, n° 26 .

(4) JAKOBSON, R., FANT, G., HALLE, M., 1963, Preliminaries to speech analysis, MIT Press, Cambrige.

(5) LAFON, J.C., 1961, Message et phonétique, Presse universitaire de France, Paris.

(6) PACHIAUDI, G., VERNET, M., FOTI, A., LAURENS, J.F., 1987, Etude de la communication parlée à l'intérieur des véhicules, Rapport INRETS n° 53.

(7) PECKELS, J.P., ROSSI, M., 1973, Le test de diagnostic par paires minimales; Adaptation au français du "Diagnostic Rhyme Test" de VOIERS, Revue d'acoustique, 27 : 245-262 .

(8) VOIERS, W.D., 1967, Performances of speech processing devices, III : diagnostic evaluation of speech intelligibility; A.F. Cambridge Research Labs, Final Report.

INTERACTION ENTRE L'ACTIVITÉ GESTUELLE ET L'ACTIVITÉ VOCALE DANS LA COMMUNICATION: ÉLÉMENTS THÉORIQUES ET MÉTHODOLOGIQUES

ISABELLE GUAÏTELLA

INSTITUT DE PHONÉTIQUE, LABORATOIRE PAROLE ET LANGAGE, URA 261 CNRS
29, AVENUE ROBERT SCHUMAN, 13621, AIX-EN-PROVENCE, FRANCE

Résumé:

L'étude des relations entre l'activité vocale et l'activité mimo-gestuelle dans la communication interpersonnelle est une problématique nouvelle, en raison, d'une part, du contexte théorique nécessaire à son émergence et, d'autre part, des besoins technologiques qu'elle suppose.

Depuis 1991 se développent au sein de l'Institut de Phonétique d'Aix-en-Provence des recherches sur le thème "geste et voix". Ces recherches envisagent plusieurs perspectives autour de la problématique suivante : peut-on parler de synchronisation entre les événements du continuum vocal et les événements gestuels accompagnateurs de la parole? Si oui, qu'est-ce que cela signifie?

Mots-clés: *communication multi-modale, geste, voix, prosodie, rythme, perception bi-modale.*

INTRODUCTION

Dans un modèle *tri-modal* de la communication (Bally, 1925, 1944; Mehrabian, 1972; Cosnier, 1988; Guaïtella 1991), on sous-entend que les échanges interpersonnels se fondent essentiellement sur l'utilisation de trois modalités de communication: le verbal, le vocal et le gestuel. Le verbal se réfère à la production de texte (lexique et construction syntaxique), il se distingue du canal vocal qui prend en compte la réalisation vocale en dehors de sa capacité à restituer le contenu verbal (chaîne segmentale, prosodie, rires, vocalisations diverses...). Le canal gestuel regroupe toute l'activité mimique et gestuelle visible.

Nous avons fait l'hypothèse que, au sein de ces trois modalités, l'activité vocale et l'activité gestuelle entretenaient des relations privilégiées (Guaïtella, 1991, 1994).

L'objet de cette communication est de préciser quelles sont les bases théoriques de ce que nous avons appelé l'approche 'geste et voix' et quels en sont les aspects méthodologiques. Par ailleurs nous illustrerons notre argumentation par l'évocation des travaux en cours dans le cadre de notre groupe de recherche à Aix en Provence.

1. D'OÙ VIENT LA PROBLÉMATIQUE ? ELÉMENTS HISTORIQUES ET THÉORIQUES

1-1. Les phonéticiens et le geste mélodique

Quelques phonéticiens ont envisagé la possibilité pour le geste accompagnateur de la parole d'être lié - par son timing, mais surtout par ses caractéristiques sémiotiques - à l'intonation de la voix.

Heese (1957) a clairement évoqué le problème en citant plusieurs cas où le geste et l'intonation produisent des trajectoires similaires: par exemple, le questionnement avec une intonation montante associée à une levée du menton, ou le poing s'abattant avec une intonation affirmative descendante... Cependant, Heese ne relève que les faits remarquables et dont l'analyse est immédiatement accessible sans besoins théoriques ou méthodologiques précis.

Bolinger (1985) propose une description plus complète de la relation geste-intonation. Il envisage de nombreux cas, et il apparaît ainsi que la relation est de nature complexe. Les mouvements tels que ceux décrits par Heese et que l'on peut qualifier de *similaires*, c'est-à-dire descriptibles par un même qualificatif ("montant" ou "descendant") pour ce qui est de l'orientation des trajectoires visuelles et auditives, ne correspondent pas à la totalité des phénomènes observés, loin de là... En effet, Bolinger relève l'existence de configurations contredisant ce schéma idyllique. Il conclut à la possibilité pour le geste et l'intonation de créer des nuances et des effets de sens par un jeu sur la mise en relation de leurs constituants. Ainsi, comme l'ironie se manifeste notamment par un décalage entre la production verbale et la production vocale, le geste et la voix pourraient également créer des effets par décalage.

Nous avons proposé une méthode d'analyse systématique de ces phénomènes par un système de notation sur des grilles (*grilles multi-paramétriques*; Guaïtella 1991). Les résultats de notre étude montrent qu'il existe effectivement des *signes intono-gestuels* de nature complexe, mais que leur diversité n'est pas due,

comme le disait Bolinger, à la concaténation de composants prosodiques et gestuels de signifiés différents, mais plutôt à l'utilisation par le locuteur de mécanismes sémiotiques divers permettant effectivement des associations voco-gestuelles différentes, mais néanmoins homogènes en raison de ce lien sémiotique. Ainsi le locuteur n'aurait pas le choix dans la coordination de tel élément gestuel avec tel ou tel élément vocal, mais - malgré la diversité - les *signes intono-gestuels* obéiraient à des règles de combinaison sémiotiques qui échappent au locuteur. Le locuteur choisirait alors la combinaison sémiotique et non pas les éléments des diverses modalités pris indépendamment (Guaïtella, 1994).

1-2. La mimique laryngée

Les phonéticiens s'intéressant au geste phonatoire - et à son interaction avec les phénomènes expressifs (Bolinger, 1946; Fonagy, 1962, 1967; Ohala 1980, 1984) - apportent des éléments de compréhension et des arguments pour une mise en relation de l'expression vocale et gestuelle. En effet, si l'on considère que l'expressivité véhiculée par la voix est dépendante d'une *mimique laryngée*, c'est-à-dire qu'elle est la trace sonore d'une activité physiologique qui, bien qu'interne, est comparable à une activité externe telle que la mimique, alors on admet qu'il peut y avoir un lien entre le visuel et l'auditif. L'absence de lien serait la manifestation d'une cassure entre l'activité gestuelle interne et externe. Ohala a montré que la mimique du sourire trouvait son origine dans l'expression vocale de l'amabilité et dans l'activité phonatoire (physiologique) correspondante. On aurait donc une activité gestuelle globale de l'individu dont une partie interne peut être rendue perceptible à l'interlocuteur par l'oralisation (la voix) et dont une partie externe est directement perceptible par la vision.

1-3. La communication non-verbale

Les chercheurs en communication non verbale et en éthologie humaine se sont - pour certains d'entre eux (Birdwhistell, 1970; Condon, 1976; Cosnier, 1988) - penché sur la double question de la *micro-analyse* du geste (selon la terminologie de Condon) et de la mise en relation du geste avec la composante vocale considérée spécifiquement par rapport au texte.

La micro-analyse de l'activité gestuelle suppose le relevé systématique des événements gestuels - par opposition à un relevé sélectionné en fonction d'un choix portant sur le type d'événements attendus - et la localisation de ces événements sur l'axe temporel - on souligne ici l'importance du timing et de la synchronisation, par opposition aux descriptions statiques du geste et de la mimique.

Dans ce courant de recherche on a donc parfois envisagé une mise en relation du gestuel avec le vocal considéré de façon relativement autonome par rapport au texte (dans le sens ou un codage des événements prosodiques a été proposé en plus de la notation du texte; Birdwhistell, Condon), mais cela s'est fait avec un décalage dans les niveaux: ainsi face à une notation plus directe des événements gestuels, le codage des informations vocales est de niveau supérieur (accents, contours intonatifs... ; voir, notamment, Condon), alors qu'on peut envisager une notation brute de la courbe. Il est bien évident toutefois que cette lacune est dépendante des moyens d'analyse dont disposent les chercheurs.

2. ACTUALITÉ DE CETTE RECHERCHE

Comme nous venons de le souligner, ce domaine est extrêmement dépendant des moyens d'analyse disponibles. Par moyens d'analyse, on peut entendre bien entendu l'infrastructure technologique, mais aussi la maturation des approches théoriques et méthodologiques.

2-1. Mutations théoriques dans l'analyse du langage

La faveur accordée actuellement à l'interdisciplinarité dans les recherches sur le langage permet à notre approche de trouver sa place. Nous considérons que la phonétique a la capacité de prendre en charge l'étude des paramètres d'expression d'ordre gestuo-visuel, comme elle le fait déjà pour les composantes gestuo-visuelles qui participent à la production des éléments segmentaux de la parole (étude articulatoire pour les phonèmes).

L'importance légitime accordée actuellement au corpus - variété des corpus, typologie... - et surtout à la production de parole spontanée amène à modifier certains concepts, considérés jusqu'alors comme centraux dans l'analyse scientifique de la parole, et qui pouvaient nuire au développement des recherches sur le lien entre le geste et la voix. Parmi ces concepts, ceux de contrôle des paramètres et de reproductibilité en situation expérimentale commencent à être un objet de discussion dans notre discipline. Un autre concept qui pose problème est celui du plan expérimental "hypothèse>vérification". Comme l'a montré Condon, les investigations dans un domaine aussi peu exploré nécessitent au contraire une démarche plus souple et plus interactive: les premières hypothèses (inévitables) seront généralement abandonnées au profit d'hypothèses provoquées par la première investigation. La démarche expérimentale devient alors plus riche en se constituant de diverses couches - éventuellement contradictoires - d'hypothèses et d'expérimentations.

2-2. Progrès technologiques

Ce domaine de recherche réunit deux types de difficultés technologiques: d'une part, la complexité liée au traitement de corpus spontané et continu, et, d'autre part, la complexité liée au traitement du mouvement, entre autres dans ces aspects de traitement du visuel. Les progrès technologiques sont donc réellement stimulants, dans le sens où ils sont une nécessité absolue et non pas simplement un moyen de faciliter le travail.

Le développement informatique permet de plus en plus le stockage et le traitement de séquences longues et

de paramètres multiples et diversifiés appréhendés simultanément. Par ailleurs, les progrès en traitement automatique et en synthèse d'images permettent l'analyse et la simulation pour vérification.

L'intérêt du traitement automatique d'images est de permettre plus d'objectivité pour le chercheur dans l'analyse des mouvements visuels - au même titre que l'analyse du signal audio permet plus d'objectivité pour l'étude de la voix qu'une analyse phonétique "à l'oreille". De plus, cette objectivité de traitement - accessible normalement par un appareillage du locuteur - peut alors être obtenue sur des documents vidéo déjà constitués (et non pas constitués dans un but expérimental).

Les méthodes d'analyse du mouvement par appareillage - quel que soit le domaine de recherche pour lequel elles ont été développées au départ - ont également progressé et présentent généralement moins de contraintes en raison de l'allègement des capteurs (méthodes ne nécessitant pas d'électrodes). On peut citer les méthodes électromagnétométriques (movetrack; Branderud, 1985) qui présentent toutefois des contraintes sur le nombre de capteurs et d'axes de déplacement potentiels. Les méthodes du type "selspot", basées sur la réflexion de la lumière par des capteurs-pastilles très légers, ne présentent pas de limite de principe au nombre de capteurs et à leurs déplacements (pour une revue bibliographique des moyens d'acquisition, voir Lallouache, 1991).

Les progrès sont donc catégorisables en progrès nécessaires - accroissement de l'objectivité dans l'analyse; possibilités de simuler; accroissement du naturel ("authenticité") dans les situations d'expérimentation par la limitation des contraintes - et progrès de confort - économie de temps et d'énergie dans la recherche.

3. EXTENSIONS ET APPLICATIONS

Le domaine d'investigation que nous essayons donc de présenter, concerne - tout comme les autres domaines de la phonétique expérimentale - plusieurs sous-domaines d'étude, aussi bien en recherche fondamentale qu'en recherche appliquée.

3-1. Recherche fondamentale

Production

L'étude des liens entre production de la voix et production du geste apporte des connaissances sur la capacité du locuteur à gérer plusieurs modalités d'expression de nature différente. Il semble que certaines dimensions vocales et gestuelles entretiennent des relations fortes en ce qui concerne leur coordination motrice. L'étude des sourcils et de la fréquence fondamentale montre que, si leur interaction est variable - en fonction du contexte, de la situation et du locuteur - lorsque cette interaction se manifeste, on assiste à une synchronisation exacte (dans la limite des techniques d'étude utilisées jusqu'à présent) (Guaïtella, Cavé & Santi, 1993; Cavé, Guaïtella & Santi, 1993; Bertrand, 1993; Scotto di Carlo & Guaïtella, étude en cours). Comparer la production de la voix et du geste devrait apporter des renseignements sur l'activité motrice de l'individu en situation de parole (Boyer, 1993). Selon notre point de vue, mettre en relation des données issues de tâches d'oralisation de texte écrit et-ou de récitation, de parole produite spontanément et de chant, aménera des éléments de connaissance, non seulement sur d'éventuelles coordinations motrices, mais également sur la relation entre activité cognitive (différente en fonction des tâches évoquées ci-dessus) et manifestation auditive et visuelle de l'activité motrice.

Perception

L'effet Mc Gurk a été défini en ce qui concerne le niveau segmental comme l'effet perceptif résultant du conflit configurationnel entre les indices visuels d'un son de parole (mouvements visibles des organes articulateurs associés à un son) et les indices auditifs (Mc Gurk & Mc Donald, 1976; Mc Donald & Mc Gurk, 1978; Cathiard, 1988-1989). Ce conflit créé artificiellement au niveau de la production par l'expérimentation, a pour conséquence du point de vue perceptif, un mixage des indices où la dimension visuelle est souvent dominante en cas d'incompatibilité. La démonstration de cet effet a remis en question l'idée d'une perception exclusivement auditive de la structure phonématique du langage oral.

Par ailleurs, le phénomène d'autosynchronie a été montré au niveau suprasegmental (Condon, 1976; Guaïtella, 1991, 1994): il correspond au fait que le locuteur coordonne la production des indices rythmiques émis à travers les paramètres gestuels et prosodiques. Ceci sous-entend que le locuteur gère la macro-organisation de son discours en produisant les indices gestuels et auditifs de façon coordonnée par rapport au timing. Il s'agit donc d'un phénomène de production dont on suppose qu'il joue un rôle du point de vue perceptif.

Deux expérimentations sont en cours visant à tester la manifestation perceptive de l'effet Mc Gurk - que nous nommons, dans nos travaux, *décalage sémiotique* - et de la perturbation de l'autosynchronie - que nous nommons *décalage rythmique* - aussi bien au niveau segmental (ce qui n'a pas été fait pour l'autosynchronie) qu'au niveau supra-segmental (Santi, expérience en cours; Andersson, Cavé, Guaïtella & Santi, expérience en cours). L'expérimentation au niveau supra-segmental suppose la prise en compte de ce qui a pu être décrit comme des signes intono-gestuels (Guaïtella 1991, 1994).

Rythme, planification du discours et interaction

L'étude de corpus interactifs suppose que le discours étudié est, d'une part, improvisé et, d'autre part, modelable en fonction de l'échange (contexte et interlocuteurs). On ne saurait donc faire de prédictions

sur la production discursive du locuteur. Par ailleurs un modèle de la communication tri-modale sous-entend que la planification du texte n'a pas de caractère *nécessairement préliminaire* à la planification du geste et de la voix. Donc, on ne peut déterminer l'organisation rythmique à partir de prédictions basées sur le texte (voir pour certains aspects de ce problème, Bolinger 1972). Nous avons montré que pour une étude de l'organisation rythmique de la parole, l'analyse du geste et de la voix permettait de remplacer les prédictions sur la structure rythmique, par des indices établis selon des critères de forme.

3-2. Recherche appliquée

Pathologie

Plusieurs recherches sont en cours dans le domaine de la pathologie. L'analyse de la synchronisation du geste et de la voix chez les malades parkinsoniens - en collaboration avec l'hôpital d'Aix - devrait permettre d'affiner les hypothèses sur la coordination motrice et, si cela s'avère justifié, de développer une échelle sur ce critère pour le diagnostic du degré de la maladie (Mignard, 1994). L'étude de cas d'aphasie transitoire (Bertrand, recherche en cours; en collaboration avec B.Cyrulnik et l'hôpital de la Seyne) apportera des données permettant également de vérifier les hypothèses générales sur la relation geste/voix.

Technologie de la parole: synthèse

L'apparition de travaux intégrant une dimension visuelle à la synthèse vocale par l'intermédiaire de la synthèse de visage parlant ("visual speech"; voir Brooke, 1992), devrait amener à prendre en compte certains éléments faciaux, dont par exemple, les mouvements de sourcils et introduire ainsi une variable expressive supplémentaire. En outre si nous parvenons à préciser les liens entre activité prosodique et faciale, on pourrait générer la prosodie et la mimique à partir d'un module de règles commun (Santi & Guaïtella, recherche en cours).

Actuellement des films peuvent être réalisés en synthèse d'images en reproduisant le mouvement capté et stocké d'un humain. Les recherches sur le geste en relation avec la voix permettront, dans un premier temps, d'affiner et de changer les réalisations ainsi obtenues - en modifiant le geste ou la voix - tout en restant dans la limite du naturel de ce qui peut être produit par un humain (du point de vue de la synchronisation et de la cohérence sémiotique) (Boyer, recherche en cours; en collaboration avec l'entreprise Gribouille à Aix-en-Provence sur le projet "20000 lieues sous les mers").

Didactique des langues

Intégrer la production gestuelle des locuteurs à l'enseignement des langues est une idée qui commence à être admise (voir les travaux de Calbris). Envisager la gestion simultanée par l'apprenant de sa production prosodique et gestuelle nous paraît intéressant. L'expérimentation que nous avons en cours sur le suédois et le français (Andersson et al.) apportera certainement des données permettant de déterminer des faits spécifiques à ces langues par rapport à d'autres phénomènes de tendance plus universelle.

CONCLUSION

Dans ce survol - certes trop rapide - nous avons essayé de montrer que l'idée d'analyser la mise en relation de l'activité vocale et de l'activité gestuelle d'un individu en train de parler est une idée qui n'est pas récente, et qui prend en grande partie sa source dans la recherche en sciences phonétiques.

Par ailleurs, nous espérons avoir souligné l'intérêt de notre approche aussi bien que le potentiel de notre groupe de recherche.

La principale conclusion, dans l'état actuel de nos expérimentations, est qu'un lien privilégié semble effectivement exister entre la production de la voix et la production des gestes accompagnateurs de la parole. Toutefois, cette conclusion ne peut être que provisoire et imprécise, en raison du caractère inachevé de plusieurs de nos expériences.

BIBLIOGRAPHIE

ANDERSSON L., 1992, "Visuella korrelat till (svenska vokalers) rundning", *Fonetik 92, papers from the sixth Swedish Phonetics conference*, ed. D. Huber, Göteborg: Chalmers University of Technology, Dep. of Information Theory, 103-6.

ANDERSSON L., CAVE C., GUAÏTELLA I. & SANTI S., "Interférences entre les canaux visuels et auditifs: comparaison interlangues", recherche en cours.

BALLY C., 1925, *Le langage et la vie*, Librairie Droz, Librairie Giard.

BALLY C., 1944, *Linguistique générale et linguistique française*, Francke Berne.

BERTRAND R., 1993, *Mise en relation de l'activité de la face avec les paramètres prosodiques et la chaîne segmentale dans un corpus d'interview télévisé*, Mémoire de DEA, Université de Provence.

BERTRAND R., thèse en cours, Université de Provence, Aix-en-Provence.

BIRDWHISTELL R., 1970, *Kinesics and context*, University of Pennsylvania Press.

BOLINGER D., 1946, "Thoughts on Yep and Nope", *American Speech*, 21, 90-5.

BOLINGER D., 1972, "Accent is predictable (if you're a mind-reader)", *Language*, 48, 633-44.

BOLINGER D., 1985, *Intonation and its parts*, Edward Arnold.

BOYER J., 1993, *Mise en relation des mouvements des bras et de la tête avec les paramètres prosodiques dans un corpus d'interview-reportage télévisé*, Mémoire de DEA, Université de Provence.

BOYER J., thèse en cours, Université de Provence, Aix-en-Provence.

BRANDERUD P., 1985, "Movetrack - A movement tracking system", *Symposium Franco-Suédois sur la parole*, GALF, Grenoble, 113-22.

BROOKE N.M., 1992, "Computer graphics synthesis of talking faces", in: Bailly, Benoit & Sawallis (eds), *Talking machines: theories, models and designs*, Elsevier, North-Holland, 505-22.

CATHIARD M.A., 1988-89, "La perception visuelle de la parole: aperçu de l'état des connaissances", *Bulletin de l'Institut de Phonétique de Grenoble*, 17-18, 109-93.

CAVÉ C., GUAÏTELLA I. & SANTI, 1993, "Fréquence fondamentale et mouvements rapides des sourcils: une étude pilote", *Travaux de l'Institut de Phonétique d'Aix*, 15.

CONDON W.S., 1976, "An analysis of behavioral organisation", *Sign Language Studies*, 13, 285-318.

COSNIER J., 1988, "Grands tours et petits tours", in: Cosnier, Gelas & Kerbrat-Orecchioni (eds), *Echanges sur la conversation*, Editions du CNRS, Lyon, 175-84.

FONAGY I., 1962, "Mimik auf glottaler Ebene" *Phonetica*, 8, 209-19.

FONAGY I., 1967, "Hörbare Mimik", *Phonetica*, 16, 25-35.

GUAÏTELLA I., 1991, *Rythme et parole: comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, Thèse de Doctorat, Université de Provence.

GUAÏTELLA I., 1994, "Mélodie du geste, mimique vocale", *Semiotica* (sous presse).

GUAÏTELLA I., CAVÉ C. & SANTI S., 1993, "Relations entre geste et voix: le cas des sourcils et de la fréquence fondamentale", *Actes du colloque Images et Langages, Multimodalité et modélisation cognitive*, Paris, 261-8.

HEESE G., 1957, "Akzente und Betleitgeärden", Sprachforum, 2, 274-85.

LALLOUACHE T., 1991, *Un poste "visage-parole" couleur. Acquisition et traitement automatique de contour des lèvres*, Thèse INP, Spécialité Image Parole, Grenoble.

Mc DONALD J. & Mc GURK H., 1978, "Visual influences on speech perception processes", *Perception & Psychophysics*, 24, 253-7.

Mc GURK H. & Mc DONALD J., 1976, "Hearing lips and seeing voices", *Nature*, 264, 746-8.

MEHRABIAN A., 1972, *Silent messages*, Wadsworth, Belmont.

MIGNARD P., 1994, DEA de neurologie (en cours), Université de Provence, Aix-en-Provence.

OHALA J.J., 1980, "The acoustic origin of the smile", *J. Acoust. Soc. Am.*, 68, 33.

OHALA J.J., 1984, "An ethological perspective on common cross-language utilization of fo of voice", *Phonetica*, 41, 1, 1-16.

SANTI S., "Décalages rythmique et sémiotique entre les canaux visuel et auditif: niveau segmental et suprasegmental", recherche en cours.

SANTI S. & GUAÏTELLA I., "Règles intono-mimiques pour la synthèse du visage parlant", recherche en cours.

SCOTTO DI CARLO N. & GUAÏTELLA I., "L'expression des émotions par la voix et la mimique dans le chant", recherche en cours.

L'OPPOSITION DE DURÉE VOCALIQUE EN ARABE : ESSAI DE TYPOLOGIE

JOMAA MOUNIR
Docteur en sciences du langage
UNIVERSITE STENDHAL
B.P. 25 – 38040 GRENOBLE CEDEX 9

Résumé

Nos études sur divers comportements de l'opposition de quantité vocalique en arabe , nous ont conduit à s'intéresser indirectement aux typologies des données quantitatives concernant l'arabe et les autres langues du monde. Dans le prolongement de cette recherche, nous essayerons donc de compléter en général ces observations typologiques avec les données les plus récentes dont nous disposons. Rappelons que parmi les différents indices sur lesquels repose la quantité, nous ne tenons compte dans cette typologie que de la variable temporelle durée acoustique des segments. Les résultats obtenus montrent que les rapports voyelles brèves/voyelles longues restent suffisamment bas, ce qui correspond au rôle que joue la durée dans le système phonologique de l'arabe.

Mots clés : typologie, quantité vocalique, durée acoustique, arabe standard et dialectal, régularité.

INTRODUCTION

Notre typologie a bénéficié essentiellement des travaux personnels et des travaux publiés par des phonéticiens et des phonologues qui se sont occupés, surtout expérimentalement de la quantité et de ses questions connexes (cf. les contributions des chercheurs arabophones in Bibliographie de l'Occident Musulman, édité en 1992 par l'Institut du Monde Arabe à Paris). Ces chercheurs s'accordent donc à classer l'arabe, ainsi que ses variétés parmi les langues et les dialectes où le contrôle de la durée vocalique et/ou consonantique est phonologique. De plus, les locuteurs de l'arabe utilisent ces différences quantitatives sur les voyelles aux quelles ils rajoutent incontestablement des différences qualitatives pour mieux assurer, par conséquent, le maintien des oppositions linguistiques (cf. sur ce point GHAZALI, 1979 pour une approche synchronique et BAKALLA pour une analyse diachronique). En ce qui nous concerne, nous essayerons d'apporter les éléments disponibles du dossier arabe sur *le comportement temporel moyen de l'opposition de quantité vocalique*. Rappelons que ces éléments sont rares et jusqu'alors dispersés. Ces types de connaissances restent cependant indispensables pour l'élaboration des règles phonétiques significatives ayant trait au phénomène de temporalité qui doit être intégré dans les modèles de reconnaissance, les systèmes de synthèse et dans les méthodes de l'enseignement de l'arabe.

1. TYPOLOGIE DU CONTRASTE DANS LES LANGUES ARABES

1.1. CRITÈRES

Nous avons retenu ici les mesures acoustiques des voyelles brèves vs. longues, et cela pour l'arabe standard, ainsi que pour ses variétés dialectales. Nous présentons ci-dessous, d'après les sources disponibles dans ce domaine, la durée des voyelles brèves, exprimée en pourcentage de celle des longues correspondantes. Pour ce faire, nous avons exprimé le rapport durée brève/longue (V/V: × 100) en pourcentage. Cet indice a été utilisé notamment par ELERT (1964), pour comparer le comportement de la quantité vocalique en suédois par rapport à d'autres langues qui en font usage. LEHISTE (1970) et NORLIN (1987) ont fait de même pour montrer l'importance de la variation de la quantité vocalique à travers les langues du monde. Dans une optique comparative, nous ne tenons compte que des mesures sur la voyelle basse [a]. Cette voyelle est la plus rencontrée dans les langues du monde. Elle est représentative de tous les systèmes vocaliques – comme dans nos données sur l'arabe (100% des 317 langues du monde citées dans UCLA Phonological Segment Inventory Data Base, possèdent la voyelle [a], MADDIESON, 1986). Si l'on croit METOUI (1989, p. 37), les voyelles /i/ et /u/ en arabe standard ont un faible pourcentage d'occurrence dans le lexique arabe, et ceci par rapport à la voyelle /a/, soit 20,8% pour /i/, 19,8% pour /u/ et 59,4% pour /a/. ABOU HAIDER (1991, p. 140) montre également que les contraintes des paires minimales signifiantes en arabe standard font que les voyelles [a] et [a:] sont largement majoritaires dans son corpus tiré d'un grand dictionnaire arabe unilingue. Cette méthode nous permet donc de tenter une comparaison des données *toutes choses égales par ailleurs*, mais pas dans tous les cas, puisque les matériaux linguistiques disponibles sont souvent très variés et sont spécifiquement structurés pour observer des phénomènes précis.

1.2. L'ARABE STANDARD

C'est sans doute l'arabe standard contemporain, appelé également arabe littéral, par opposition à l'arabe classique et à l'arabe dialectal, qui a suscité le plus d'intérêt dans les travaux des chercheurs, comme on peut le constater ci-après.

– PORT *et al.* (1980, p. 239) montrent que le rapport de la voyelle brève par rapport à la longue, en arabe standard prononcé par des locuteurs égyptiens, irakiens et koweïtiens, est de 39%. En débit conversationnel, la durée de la voyelle [a] suivie de [t] est en moyenne de 60 ms et de 155 ms pour la longue.

– BELKAID (1984, p. 228-230) a montré que la durée moyenne des voyelles cardinales brèves [i], [u] et [a], dans la première syllabe accentuée de l'arabe standard prononcé par un locuteur à substrat dialectal tunisien, est supérieure ou égale à la double de leurs correspondantes longues. La voyelle [i] précédant une consonne nasale [n], avec 60 ms, représente 40% de la durée de la longue correspondante ([i:] étant de 150 ms). La voyelle brève [u], suivie d'une occlusive [k] ne représente que 50% de la voyelle longue correspondante [u:], soit respectivement 70 ms et 140 ms. Quant à la voyelle brève [a] suivie également d'une consonne nasale [n], avec 75 ms, elle représente un rapport V/V: comparable à celui de la voyelle [u], soit 38%. La durée de la longue correspondante [a:] étant de 200 ms.

– ALIOUA (1987) a également étudié l'organisation temporelle des voyelles et des diphtongues de l'arabe standard, et ceci pour trois locuteurs d'origine marocaine. Les résultats temporels obtenus sur les voyelles révèlent une remarquable homogénéité entres les trois sujets. Pour le locuteur n°1, ALIOUA (1987, p. 262) trouve que la voyelle [a] suivie d'une nasale dans des mots dissyllabiques représente 39% de la durée de la longue correspondante, soit 70 ms contre 182 ms. Concernant locuteur n° 2, la voyelle brève [a], avec une durée de 67 ms représente 41% de la durée de la longue ([a:] étant de 164 ms). Le locuteur n°3 représente, quant à lui, un rapport brève/longue pour les voyelles [a] et [a:] de 33% (soit 58 ms contre 176 ms).

– ABOU HAIDER (1991) a étudié la variabilité et l'invariance du système vocalique de l'arabe standard. Trois locuteurs arabophones d'origine dialectale différentes ont été demandés de produire les syllabes arabes CVC et CVCV dans des mots isolés. Les phonèmes consonantiques choisis sont /t, b, d, m, n, f, s, ʃ, z, l/. Seules les moyennes sur les voyelles [a] vs. [a:] dans des paires minimales ont été retenues par l'auteur. ABOU HAIDER (1991, p. 106-107) trouve donc un rapport V/V: de 56% pour son locuteur A (soit 122 ms contre 219 ms). Par contre, les différences quantitatives [a] vs. [a:] sont mieux exprimées pour ses locuteurs B (avec 127 ms contre 383 ms) et C (avec 109 ms contre 335 ms), soit un rapport V/V: identique de 33%.

– GHAZALI et BRAHAM (1992, leur tableau n° 3, p. 22) ont observé le comportement de la quantité vocalique en arabe standard, à travers deux extrêmes intrinsèques, en l'occurrence [u] et [a] et selon deux types de syllabes : ouverte vs. fermée. Ces derniers ont demandé à 3 sujets dont le substrat dialectal est tunisien de produire des dissyllabes, avec un accent sur la première syllabe, et ceci dans deux conditions de débits, normale vs. rapide et selon deux styles de prononciation, ordinaire vs. pédagogique. Nous ne citons ici que les mesures fournies sur l'opposition [a] vs. [a:], dans un contexte simple [d] et double [dd] postvocaliques, et ceci uniquement pour deux locuteurs. Ces oppositions illustrent bien des véritables paires minimales en arabe standard insérées dans des phrases porteuses. En arabe standard, la durée des brèves [a] est de 55 ms en moyenne pour le locuteur SBR. Les longues sont de 109 ms. Le rapport brève/longue est donc de 50%. Rappelons que la brève [a] suivie d'une occlusive double est de 48 ms, soit 44%. Le locuteur A.B. présente un rapport V/V: de 49% (soit 66 ms et 136 ms respectivement). Notons également que la brève [a] suivie d'une occlusive double est de 61 ms pour ce locuteur, soit 45%. Les deux locuteurs ont donc un comportement de durée similaire.

– ALGHAMDI (1990) trouve pour les voyelles [a] et [a:] un rapport de 52%, soit respectivement 88 ms et 151 ms.

– Nous avons montré dans un travail très récent sur l'opposition de quantité en arabe standard comme indice de gémination et/ou de dédoublement consonantique (JOMAA, à paraître) que la durée de la voyelle brève [a] suivie de [t] (avec 64 ms de durée en moyenne) représente 36% de la durée de la longue (avec 176 ms). La voyelle brève [a] suivie de [tt] est de 40 ms, soit un rapport V/V: de 23%.

Il nous semble maintenant intéressant d'analyser sommairement le comportement de la quantité vocalique dans les variétés de l'arabe, considérées comme des langues essentiellement à traditions orales et des outils de communication vernaculaire.

1.3. LES VARIETES DE L'ARABE

Arabe tunisien : pour ce qui est du dialecte tunisien et uniquement des voyelles dites basses, les résultats temporels sont à peu près identiques. Nous avons montré que la durée de la voyelle brève [a] suivie de [t] (avec 67 ms de durée en moyenne) représente environ 59% de la durée de la longue (avec 113 ms), en débit conversationnel (JOMAA, 1987). Concernant les voyelles [ɛ] et [ɛ:] suivies d'une constrictive simple [z] et double [zz], le rapport V/V: est de 57%, soit 75 ms contre 132 ms JOMAA (1991, p. 147). GHAZALI (1979, p. 211) a montré que la voyelle postérieure [ɑ] suivie de la consonne [b], avec une durée moyenne de 75 ms représente 60% de la durée de la correspondante longue ([ɑ:] étant de 125 ms).

une variabilité *interdialectale* de l'opposition de quantité vocalique. Ces différences de comportement peuvent par la même nous renseigner sur la spécificité de chaque dialecte arabe, comme cela semble se produire en arabe marocain où l'opposition de durée consonantique est en effet plus manifeste que celle de la durée vocalique (RHARDISSE, 1989).

2.2. SITUATION DES TYPES DE CONTRASTES : LES LANGUES ARABES ET LES AUTRES

Pour procéder à des comparaisons globales du comportement de la quantité vocalique en arabe avec les différentes langues du monde que nous avons recensées, nous reprenons la typologie établie dans JOMAA (1991). Dans cette typologie (cf. figure 3 reflètant une variabilité *interlangue* (apparentée ou non apparentée), les langues sont classées en 5 groupes :

1- *Langues à très fort contraste* : ce groupe n'est représenté que par une langue : le telgou dont le rapport V/V: est autour de 35%.

2 - *Langues à fort contraste* : concernant ces langues, le rapport V/V: est toujours inférieur ou égale à 50%. C'est notamment le cas du danois, du tamil, du tchèque, du malayalam, du breton, du finnois, du luli sami, du luxembourgeois, du wolof, du cantonnais, du francoprovençal et enfin du néerlandais.

3 - *Langues à contraste standard* : dans ce groupe, le rapport V/V: se situe autour de 50%-60%. C'est le cas du yavapai, du haoussa, du mbay, du hongrois, de l'estonien, du thaï, du nar, du norvégien et de l'islandais.

4 - *Langues à faible contraste* : pour ces langues, les voyelles brèves représentent entre 80% et 65% des longues correspondantes. A cette classe, appartiennent le suédois, l'anglais, le français, le béjonde, l'italien, le sar et le serbocroate.

5 - *Langues à très faible contraste* : ce sont les langues pour lesquelles le rapport V/V: est autour de 90%. Dans ce groupe, nous ne trouvons que l'allemand et le flamand.

Dans cette classification générale (cf. Figure 3), nous pouvons constater que l'arabe standard n'apparaît généralement que dans les deux premiers groupes où l'opposition de quantité vocalique est optimale. Les dialectes arabes tels que le koweïtien, le syrien et le libanais apparaissent souvent soit dans le deuxième groupe où les langues sont majoritaires, soit dans le troisième groupe qui représente l'arabe saoudien, tunisien, égyptien et jordanien. L'arabe marocain, quant à lui, appartient au quatrième groupe, avec un rapport V/V: qui est de 77%et qui est donc comparable à celui du suédois.

3. CONCLUSION

Les éléments typologiques que nous avons récencés sur le comportement de l'oppositions de durée vocalique nous ont permis de confirmer que la durée phonétique des voyelles en arabe correspond parfaitement à saquantité phonologique : *une voyelle longue présente une durée égale au moins au double de celle de la brève correspondante*, ce qui nous donne un rapport 2:1. C'est notamment les cas de l'arabe standard et de certains dialectes qui manifestent selon cette typologie une régularité temporelle remarquable du contrôle des oppositions linguistiques, avec toutefois des stratégies du contrôle du temps propre à chaque dialecte. Par conséquent, nous devons tenir compte de ces différences temporelles entre les voyelles brèves et les voyelles longues, afin d'assurer par exemple une certaine naturalité de la synthèse du système sonore de l'arabe, et plus particulièrement de son vocalisme. Il reste encore à déterminer si le corrélat acoustique de durée vocalique est à lui seul une condition nécessaire et suffisante pour marquer les oppositions linguistiques pour des dialectes arabes où la différence de durée est souvent accompagnée par une différence de timbre (cf. GHAZALI, 1979 et NORLIN, 1987).

Enfin, il évident que notre typologie n'a pas été menée de façon systématique et exhaustive sur la totalité du comportement vocalique à travers toutes les variétés de l'arabe. La faisabilité de cette entreprise n'est possible qu'en commençant par :

1- élargir notre typologie en explorant le domaine consonantique parce que l'on sait que les caractéristiques temporelles des voyelles brèves vs. longues se révèlent entretenir généralement des relations d'interdépendances avec les consonnes adjacentes. Nous ne pouvons ici qu'énumérer les champs recherches sur l'organisation temporelle où les connaissances seront indispensables. Il s'agit donc d'étudier en particulier l'influence du phénomène de l'effet du voisement consonantique sur l'opposition de durées des voyelles (cf. sur ce point FLEGE, 1979 ; PORT *et al.* 1980 et plus récemment HUSSAIN, 1985 ; ALGHAMDI, 1990 et ABOU HAIDER, 1991). Qu'en est-il également du comportement de la durée en relation avec l'emphase et la gémination distinctive (cf. également les premiers résultats in BONNOT, 1979 ; NORLIN, 1987 ; IRISOUSSI, 1981 ; GHAZALI et BRAHAM, 1992 et JOMAA (à paraître). Il serait aussi intéressant de voir si nos variétés de l'arabe résistent ou non aux variations temporelles du débit ou vitesse d'élocution (cf. JOMAA, 1991 où nous disposons de quelques indications sur la résistivité du type de contraste à la vitesse d'élocution en arabe et en d'autres langues, et plus récemment GHAZALI et BRAHAM, 1992).

2- généraliser les études sur l'organisation temporelle de la quantité à l'ensemble du système vocalique et consonantique de l'arabe en tenant compte des aires linguistiques du Monde arabe. Si les études sur les descriptions phonologiques sont relativement abondantes dans la littérature, les études expérimentales, surtout articulatori-acoustiques correpondantes ne sont qu'à leurs débuts.

Arabe marocain : en arabe marocain, la durée moyenne de la voyelle [a] suivie d'une occlusive sonore [d] est de 108 ms, contre 139 ms pour la longue correspondante en débit normal, ce qui nous donne un rapport brève/longue très élevé de 77% (RHARDISSE, 1989, p. 29).

Arabe égyptien : en arabe égyptien (NORLIN, 1987, p. 78) et pour un débit normal, la durée des voyelles brèves [i, a, u] (avec 85 ms) représente en moyenne 56% des voyelles longues correspondantes (avec 152 ms).

Nous ne tenons compte que des mesures sur la voyelle [a] suivie de [t]. Le rapport brève/longue pour la classe [a] est de 59% (soit 89 ms contre 151 ms). Ce résultat correspond exactement à celui obtenu pour l'arabe tunisien.

Arabe libanais : pour l'arabe libanais, le rapport de la brève à la longue est autour de 50% (OBRECHT, 1968, cité par NORLIN, 1987, p. 79). Il est également intéressant de citer l'ancien résultat de DURAND (1946, p. 61) sur l'opposition de quantité vocalique dans ce dialecte. Elle a montré que la voyelle brève [a] dans des monosyllabes dure 110 ms contre 220 ms pour la voyelle longue correspondante, ce qui nous donne un rapport V/V: de 50%. Dans des mots dissyllabiques, la voyelle brève [a] mesure 85 ms contre 140 ms pour la longue correspondante, donnant un rapport V/V: de 61%. Ces résultats confirment généralement ceux d'OBRECHT (1968).

Arabe jordanien : en arabe jordanien (MITLEB, 1984, p. 231), la voyelle brève [a] suivie de [t, d] présente en moyenne une durée de 90 ms. La longue [a:] est de 145 ms suivie de [tt, dd]. Le rapport brève/longue est donc de 62%.

Arabe irakien : AL-ANI (1970) montre pour l'arabe irakien, que le rapport de la brève à la longue, dans un contexte de mots monosyllabiques isolés, est de l'ordre de 50% (soit 300 ms et 600 ms respectivement).

Arabe saoudien : FLEGE (1981, p. 127) cite le résultat qu'il a obtenu sur l'opposition de la quantité vocalique en arabe saoudien (FLEGE, 1979). Il a montré que la voyelle brève [a], dans un contexte d'occlusives, est de 87 ms ; la longue correspondante est de 177 ms, ce qui nous donne un rapport V/V: de 55%. Ce résultat est corroboré par HUSSAIN (1985, cité par ALGHAMDI, 1992, p. 82) qui trouve également un rapport V/V: de 55% pour les voyelles [a] et [a:], soit respectivement 112 ms 203 ms.

Arabe koweïtien : en arabe koweïtien, en débit conversationnel, la durée des brèves [a] suivies d'une emphatique simple [ṣ] est de 73 ms en moyenne. Les longues sont de 152 ms. Le rapport brève/longue est donc de 48% (AL-DOSSARI, 1989, p. 36). Rappelons que la brève [a] suivie d'une emphatique double est de 88 ms, soit 58%.

Arabe syrien : IRISOUSSI (1981, p. 41 et 77) a étudié, entre autres, la quantité vocalique dans le parler arabe de Damas (Syrie), et ceci pour trois locuteurs. Ce dernier montre que la durée moyenne des voyelles brèves inaccentuées /i, e, a, o, u/ est de 49 ms ; les longues correspondantes /i:, e:, a, o:, u:/ est de 119 ms, ce qui nous donne un rapport V/V: de 41%. En position accentuée, les voyelles brèves ne représentent, quant à elles, que 50% de la durée des longues correspondantes, soit respectivement 63 ms et 126 ms. Nous ne tenons compte que du rapport brève/longue pour la classe [a] suivie de l'occlusive [t] qui est de 49% (soit 55 ms contre 151 ms). Pour cette même classephonologique, le rapport V/V: passe de 49% à 53% en contexte d'une occlusive sonore [d], soit 58 ms contre 110 ms.

2. RESULTATS

2.1. SITUATION INTRALANGUE DES TYPES DE CONTRASTES : L'ARABE STANDARD ET LES VARIÉTÉS DIALECTALES

La figure 1 résume le comportement général de la quantité vocalique en arabe standard. Le rapport voyelles brèves/voyelles longue varie entre 33% et 56%, avec une moyenne 42% pour l'ensemble des données concernant l'arabe standard. Ce rapport dépasse rarement le 50%. Nous constatons donc que l'arabe standard présente *un très fort contraste* entre les voyelles brèves et les voyelles longues (c. infra). Certains résultats montrent même qu'une voyelle longue en arabe standard dure parfois trois fois plus que la brève (cf. supra). Notons également que les valeurs absolues obtenues révèlent une remarquable homogénéité à l'intérieure même de chaque classe phonologique, à l'exception bien entendu des résultats obtenus par ABOU HAIDER (1991) où les valeurs sont très élevées, dû incontestablement au contexte des mots isolés qu'elle a utilisé (cf. tableau 1).

La figure 2 nous donne les rapports voyelles brèves vs.voyelles longues en pourcentage dans les variétés de l'arabe que nous avons citées. Nous remarquons que tous tous les dialectes manifestent des différences suffisantes de durée vocalique, à l'exception de l'arabe marocain. Les variétés de l'arabe, tels que le koweïtien (48%), le syrien (49%) et le libanais (50%) présentent donc des résultats convergeants avec ceux de l'arabe standard offrant un bon contraste, avec un rapport V/V: toujours inférieur ou égale à 50%. Les autres variétés s'écartent légèrement de l'arabe standard. Cependant, elles offrent un contraste dit *standard*, avec un rapport V/V: compris entre 55 et 60% (cf. tableau 2).

Nous constatons donc que les rapports restent dans l'ensemble suffisamment bas, à l'exception bien entendu de l'arabe marocain qui représente une extrême, avec un rapport V/V: de 77%. Ce résultat général correspond au rôle que joue la durée dans le système phonologique de l'arabe. Mais, nous assistons généralement pour l'arabe à

4. BIBLIOGRAPHIE

ABOU HAIDER (1991) : Variabilité et invariance du système vocalique de l'arabe standard. Thèse de Doctorat. Université de Franche-Comté. Bésançon.

AL-DOSSARI (1989) : Le phasage des gestes mandibulaires vocaliques et consonantiques en arabe koweïtien. Mémoire de D.E.A. de Sciences du langage, Grenoble III.

ALGHAMDI M.M. (1990) : Analysis, synthesis and perception of voicing in Arabic. Unpublished Ph. D. Dissertation. University of Reading.

ALIOUA A. (1987) : Etude phonétique et morpho-phonologique des voyelles et des diphtongues de l'arabe littéral moderne. Thèse de Doctorat. Strasbourg II.

BAKALLA M. H. (1983) : The study of vowel quality and quantity by early Arabs and Muslim phoneticians. Proceedings of the 10th ICPS, Utrecht.

BELKAID Y. (1984) : Les voyelles de l'arabe littéraire moderne : analyse spectrographique. T.I.P.S. 16, 217-240.

BONNOT J.F.P. (1979) : Etude expérimentale de certains aspects de la gémination et de l'emphase en arabe. T.I.P.S. 11, 109-118.

DURAND M. (1946) : Voyelles longues et voyelles brèves : essai sur la nature de la quantité vocalique. Paris, Klincksieck.

ELERT C.C. (1964) : Phonological quantity in Swedish. Stockholm-Göteborg-Uppsala.

FLEGE J. E. (1979) : Temporal correlates of [voice] in Arabic-accented English. J.A.S.A., 65, suppl. 1, S32.

FLEGE J.E. (1981) : Cross-language phonetic interference : Arabic to English. Language and Speech 24, 125-146.

GHAZALI S. (1979) : Du statut des voyelles en arabe. Analyse Théorie 2/3, 199-219.

GHAZALI S. et BRAHAM A. (1992) : Voyelles longues et voyelles brèves en arabe standard : organisation temporelle. 19èmes J.E.P. du GCP de la SFA, 19-23.

HUSSAIN A. A. (1985) : An experimental investigation of somme aspects of the sound system of the Gulf Arabic with special reference to duration. Unpublished Ph. D. Dissertation. University of Essex.

IRISOUSSII. (1981) : Analyse acoustique et phonétique des consonnes de l'arabe dans le parler de Damas. Thèse de 3ème Cycle, Université de Provence.

JOMAA M. (1987) : Etude sur l'organisation temporelle de l'opposition de quantité vocalique en arabe tunisien. Sa résistivité aux variations de la vitesse d'élocution. Mémoire de D.E.A de Sciences du langage, Grenoble III.

JOMAA M. (1991) : Organisation temporelle et acoustique et articulatoire de la quantité en arabe tunisien. Thèse de Doctorat, Université Stendhal, Grenoble III.

JOMAA M. (à paraître) : Some cues to gemination in Arabic. XIII th ICPS, Stockholm, 1995.

LEHISTE I. (1970) : Suprasegmentals. MIT, Cambridge, Massachusetts.

MADDIESON I. (1986) : Patterns of sounds. 2nd ed., Cambridge University Press, Cambridge.

METOUI M. (1989) : Contribution à la phonétique arabe : étude acoustique et articulatoire des voyelles du parler de Tunis. Thèse de doctorat, Université de Trêves, Schulz-Kirchner Verlag, Idstein.

MITLEB F. M. (1984) : Vowel length in Arabic and English : A spectrographic test. J. of Phonetics 12, 229-235.

NORLIN K. (1987) : A phonetic study of emphasis and vowels in Egyptian Arabic. Working Papers 30, Lund University, Department of Linguistics, 1-119.

OBRECHT D.H. (1968) : Effects of the second formant on the perception of velarization consonants in Arabic. Mouton, The Hague-Paris.

PORT R.F. AL-ANI S. & MAEDA S. (1980) : Temporal compensation and universal phonetics. Phonetica 37, 235-252.

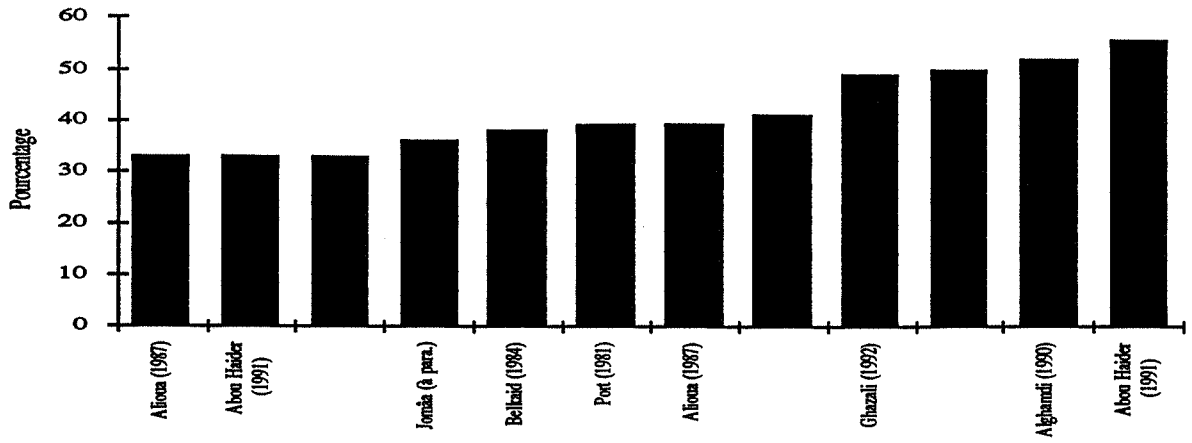
RHARDISSE N. (1989) : La résistivité de la quantité vocalique et consonantique de l'arabe marocain aux variations de la vitesse d'élocution. T.E.R. de Sc. du langage, Grenoble III.

Tableau 1 : Rapports V/V: pour les voyelles [a] et valeurs moyennes absolues correspondantes en arabe standard.

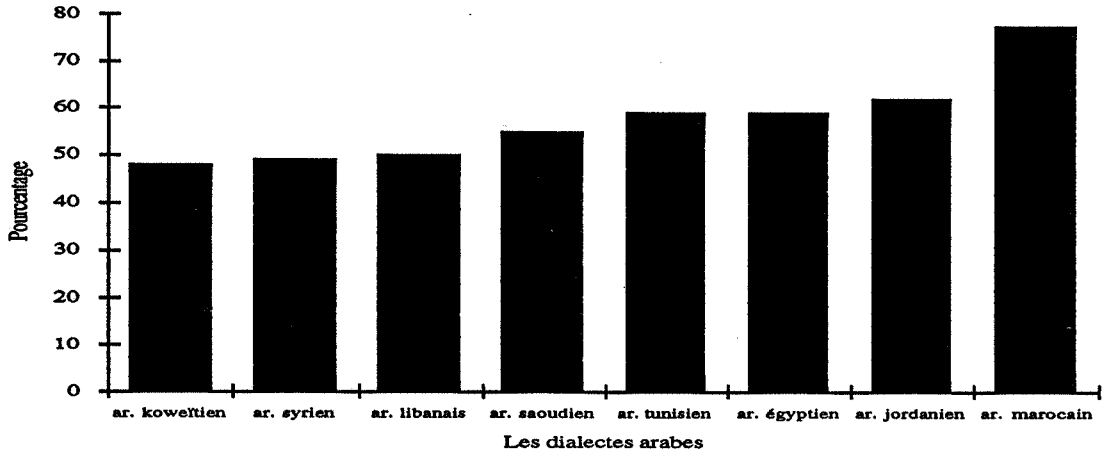
Références	Durées moyennes des voyelles en millisecondes		
	%	Voyelles brèves	voyelles longues
Port *et al.* (1980)	39	60	155
Belkaïd (1984)	38	75	200
Alioua (1987)	39	70	182
	41	67	164
	33	58	176
AbouHaider (1991)	56	122	219
	33	127	383
	33	109	335
Ghazali et Braham (1992)	50	55	109
	49	66	136
Jomâa (à paraître)	36	64	176
Al Ghamdi (1990)	52	88	151

Tableau 2 : Rapports V/V: pour les voyelles [a] et valeurs moyennes absolues correspondantes dans les variétés de l'arabe.

Variétés de l'arabe et références	Durées moyennes des voyelles en millisecondes		
	%	Voyelles brèves	voyelles longues
arabe tunisien			
JOMAA (1987)	59	67	113
JOMAA (1991)	57	75	132
Ghazali (1979)	60	75	125
arabe marocain			
Rhardisse (1989)	77	108	139
arabe égyptien			
Norlin (1987)	59	89	151
arabe libanais			
Obrecht (1968)	50	–	–
Durand (1946)	61	85	140
arabe jordanien			
Mitleb (1984)	62	90	145
arabe irakien			
Al-Ani (1970)	50	300	600
arabe saoudien			
Flege (1979)	55	87	177
Husain (1985)	55	112	203
arabe syrien			
Irisoussi (1981)	50	63	126



Figue 1 : Durée des voyelles brèves, exprimées en pourcentage des longues correspondantes en arabe standard (cf. texte)



Figue 2 : Durée des voyelles brèves, exprimées en pourcentage des longues correspondantes, pour 8 variétés de l'arabe. Dans l'ordre : arabe koweïtien (48%), arabe syrien (49%), arabe libanais (50%), arabe saoudien (55%), arabe tunisien (59%), arabe égyptien (59%), arabe jordanien (62%) et arabe marocain (77%) (cf., texte).

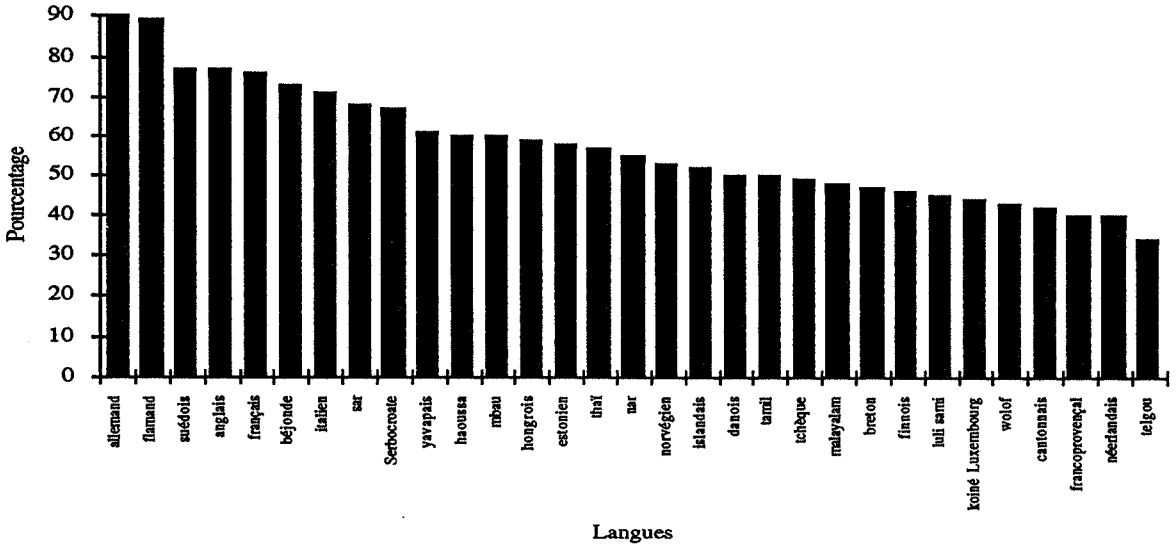


Figure 3 : Durée des voyelles brèves, exprimées en pourcentage des longues correspondantes, pour 31 langues. Dans l'ordre : telgou (34%), néerlandais (40%), francoprovençal (40%), cantonnais (42%), wolof (43%), luli sami (45%), koiné de Luxembourg (44%), finnois (46%), breton (47%), malayalam (48%), tchèque (49%), danois (50%), tamil (50%), islandais (52%), norvégien (53%), nar (55%), thaï (57%), estonien (58%), hongrois (59%), haoussa (60%), mbay (60%), yavapai (61%), serbocroate (67%), sar (68%), italien (71%), béjonde (73%), français (76%), suédois (77%), anglais (77%), flamand (89%) et l'allemand (90%), (cf. texte).

TESTS DE HAUTEUR ET PERCEPTION DE L'ACCENT EN ANGLAIS PAR DES FRANCOPHONES : UNE PREMIÈRE APPROCHE

Alain DIANA

IUT 1 , Université Joseph Fourrier
Domaine Universitaire, BP 67
38402 St Martin d'Hères cedex
et
ICP, URA 368, Université Stendhal / INPG
D.U., BP 25
38040 Grenoble cedex

Résumé

Ce travail présente les résultats d'une expérience sur la manière dont les francophones perçoivent accent et intonation en anglais. La première partie de l'expérience portait sur la détermination de la hauteur de sons non linguistiques. La deuxième partie consistait à délimiter les phrases et déterminer les syllabes accentuées d'un texte anglais radiophonique.
Les résultats donnent un pourcentage de réussite satisfaisant pour la plupart des tests, ce qui amène à formuler l'hypothèse que la perception des éléments prosodiques en L2 n'est peut-être pas très différente de la perception de ces éléments en L1. Il conviendrait de mettre au point de nouvelles expériences associant des tests élaborés à partir d'éléments non linguistiques et des tests portant sur le langage parlé, à la fois en L1 et en L2.

Mots clés : *accentuation, tests de hauteur, perception de l'anglais en L2 par des francophones,*

I) INTRODUCTION:

Peu d'études expérimentales ont été consacrées à la perception des éléments prosodiques dans une langue étrangère (L2), contrairement à celles sur la langue maternelle (L1).

Sauer [1975], par exemple, a montré qu'il existait une forte corrélation entre les scores de musicalité et les tests d'identification / discrimination en langues étrangères pour de jeunes élèves, mais ne mentionne pas l'intonation;

de Bot [1980] s'interroge sur l'utilité d'un entraînement à la discrimination auditive comme préalable à tout programme d'enseignement des langues, car (p 58): "there is not enough knowledge available about the aspects of intonation in languages which is perceptually relevant ['t Hart and Cohen, 1973] to be able to formulate such a precourse."

T. Scott [1985] se demande (p 118) "pourquoi les français ont des difficultés à reproduire les patrons de proéminence en anglais, alors qu'ils sont capables de les reconnaître aussi bien que les anglophones." Cette affirmation reste à démontrer dans la mesure où le corpus utilisé par Scott était trop réduit pour en tirer des conclusions définitives.

La présente étude tente de cerner certaines des difficultés spécifiques qu'éprouvent les francophones à percevoir certains schémas accentuels ou contours mélodiques de l'anglais. Elle se propose de rapprocher l'aptitude à repérer des changements de hauteur pour des fréquences non-linguistiques de celle qui consiste à délimiter avec précision les unités syntaxiques dans une L2, qui est liée à la perception des éléments prosodiques. En effet, en anglais, du point de vue de la communication tout au moins (conversation entre 2 interlocuteurs par exemple), il est important de pouvoir discriminer entre chutes hautes ou basses et montées hautes ou basses [O'Connor et Arnold, 1973].

II) EXPÉRIENCES:

A) Tests d'écoute (annexe I, 1):

Ces tests ont été proposés à des étudiants non-spécialistes d'anglais n'ayant aucune notion de prosodie, d'accentuation ou même de phonétique en général.

Les stimuli (sons purs : fréquences générées à l'aide d'un générateur de signal carré) ont été enregistrés (sur un magnétophone Revox 19 cm/s à 0 db) à l'Institut de la Communication Parlée de Grenoble (Université Stendhal). La durée des stimuli était d'une seconde , un intervalle de 3 secondes séparant la présentation de deux stimuli.

Dans les expériences A et B (annexe I, 1) les sujets devaient comparer 2, puis 3 fréquences distinctes, séparées par des intervalles de 500 ms. Ils devaient déterminer, dans l'expérience A la position du second son entendu par rapport au premier, dans l'expérience B lequel des sons entendus était le plus grave.

Dans les expériences C et D (annexe I, 1), les sujets devaient évaluer le sens de la pente d'une évolution continue de fréquences (monte, descend ou reste stable).

Le choix des fréquences entre 100 Hz et 200 Hz (cf. annexe I, 1) correspond en moyenne au fondamental des voix homme / femme, les différences testées étant de l'ordre du demi-ton, ou du quart de ton, donc inférieures à ce qui est considéré comme le minimum pour les différences significatives du point de vue de la communication. 't Hart [1981] considère que: "not less than 3 semitones...play a communicative role in speech", alors que pour Rietveld et Gussenhoven [1985] le minimum est de 3/4 de ton.

B) Tests de délimitation: (annexe II)

Le texte proposé était une publicité radiophonique (enregistrée à Londres sur "LBC radio"), dite par une voix féminine ayant un accent anglais standard (R.P.) avec des variations prosodiques relativement importantes, d'après les anglophones du groupe de contrôle. (Pour l'étude du Fo et une comparaison des résultats anglophones / francophones, voir Diana, CELDA, à paraître).

Ce test comportait deux phases:

a) Première écoute: délimiter les phrases du texte. C'est un exercice directement compréhensible pour des non-spécialistes qui permet de juger en première approximation si les indices qui servent à délimiter les phrases (groupes de souffle, pauses, intonation) sont globalement bien perçus.

b) Deuxième écoute: les sujets devaient souligner sur leur feuille de réponse les mots ou syllabes accentués dans chaque phrase, ligne après ligne, avec 10 secondes de pause à la fin de chaque ligne, dans un texte sans ponctuation pour éviter d'éventuelles interférences dues à la position syntaxique des éléments accentués.

C) Sujets et conditions d'écoute:

Les sujets étaient 51 étudiants en électronique de l'IUT 1 de Grenoble: 36 étudiants de 2ème année d'un bon niveau général en anglais, ainsi que 15 étudiants de 1ère année, de niveau moyen. Par ailleurs, trois enseignants anglophones ont constitué un groupe de contrôle.

Les conditions d'écoute étaient les suivantes:

- pour le test d'écoute: un seul passage pour chacun des tests; chaque sujet a écouté les enregistrements individuellement dans un casque haute-fidélité au laboratoire de langues de l'IUT 1 via un magnétophone Revox à bandes (bande passante de 40 à 20.000 Hz). Les sujets avaient devant eux une feuille de réponse comportant uniquement l'énoncé ainsi que le nombre de stimuli pour chaque test.
- pour le test de délimitation: chaque sujet a écouté les enregistrements individuellement dans un casque haute-fidélité au laboratoire de langues de l'IUT via un magnétophone Revox à cassettes (bande passante de 50 à 13.000 Hz). Les sujets disposaient d'une feuille de réponse comportant 2 versions orthographiques non-ponctuées du passage enregistré. Ils ont écouté ce passage 2 fois, une fois pour délimiter les phrases, une autre fois pour souligner les mots ou syllabes accentuées (annexe II).

III) RÉSULTATS:

1) TEST D'ÉCOUTE: (annexe I, 2)

Test A)

Plus de 75 % de bonnes réponses au minimum. Si les résultats sont sensiblement moins bons pour les sons n°2 et 6, il n'y a pas lieu de s'en étonner, les différences jugées étant de l'ordre du quart de ton (voir annexe I, 1). Par contre, pour le n°3 où les réponses sont moins bonnes que pour le n°5, alors que l'intervalle est le même, l'ordre de présentation a dû jouer, dans la mesure où les sujets n'avaient pas encore de référence pour deux sons identiques.

Test B)

Pour les n° 1 et 4: les résultats sont moins bons (60% et 50% de réponses correctes seulement). Cela confirme le rôle de la mémoire auditive, car le son le plus grave est le premier des trois et il est plus difficile d'effectuer une comparaison entre le premier et le troisième. En ce qui concerne le n°4, il n'y a qu'un quart de ton de différence sur une fréquence plus aiguë entre le premier et le dernier, ce qui est difficile à discriminer. D'où 20% de réponses considérant qu'ils sont au même niveau.

Test C)

La chute d'un demi-ton qui est pratiquement identique pour les n° 3 et 5 est bien perçue. On remarque que pour le n°4, la chute de 2 tons et demi, qui est du même ordre que ce qu'on nomme "high fall" en anglais [O'Connor et Arnold, 1973], est nettement moins bien perçue, et est confondue avec une montée pour 15 % des réponses environ. Or, pour le n°2, la chute d'un quart de ton, bien que plus dans l'aigu, est légèrement mieux perçue. En extrapolant pour la parole, on peut penser qu'un entraînement à distinguer par exemple "high fall" et "high rise" en anglais, n'est peut-être pas superflu. Pour les montées, le n°6 est moins bien perçu que le n°1, ce qui est logique: l'écart en 6 n'est que d'un quart de ton, par rapport à un demi-ton en 1.

Test D)

Ce test s'inspire de Buyschaert [1985]. A propos du choix des courbes, à l'exception du n°2 dont la courbe n'a aucun rapport avec l'intonation anglaise, mais concerne la perception de la différence des niveaux de départ et d'arrivée, les trois autres stimuli se rapprochent de l'évolution du Fo pour des schémas intonatifs de l'anglais. Les n°1 et n°3, qui sont moins bien perçus, se rapprochent respectivement des "fall-rise" et "rise-fall" en anglais [O'Connor et Arnold, 1973], ce qui pourrait confirmer que ces contours sont difficiles à percevoir [A. Raphaël, 1971]. Au contraire, les résultats sont bons pour le n°4 qui est une montée nettement marquée (combinaison des "low rise" / "high rise" de O'Connor et Arnold, 1973).

Dessiner un son est un exercice inhabituel qui a peut-être dérouté certains sujets et qui gagnerait en précision à être remplacé par un QCM, en proposant trois courbes différentes et en demandant aux sujets de déterminer laquelle se rapproche le plus de la courbe entendue.

2) Perspectives:

Il conviendrait de concevoir un test qui synthétise des courbes extraites de phrases d'un corpus en anglais, ce qui aurait le mérite de mieux maitriser les paramètres essentiels de fréquence, intensité et durée. Il resterait à établir le lien entre perception et (re)production de schémas intonatifs particuliers (avec ou sans visualisation des courbes intonatives).

2) TEST DE DÉLIMITATION:

a) Délimiter les phrases (annexe II):

On constate assez peu d'erreurs sur le placement des limites de phrases (6 pour 22 sujets), mais on trouve beaucoup moins de délimitations intermédiaires que ce que l'on obtiendrait avec une analyse des pauses.

L'absence de délimitation entre "tableware" et "it's" (lignes 3/4), est certainement due à la non mise en valeur de "tableware", ce qui produit une suite de syllabes inaccentuées ou peu accentuées et rend ainsi la délimitation plus malaisée. Le même phénomène se retrouve pour d'autres erreurs telles que: "patterns yet/it's"(l 6). On le retrouve combiné à un "High Fall" à trois reprises: "good job/it's" (l 3), "You/can't" (l 8), "available/at Selfridges"(l 9). Dans ces trois cas, le "high fall" semble être perçu comme étant le début d'une phrase. On peut aussi faire le rapprochement avec l'exemple "patterns yet/it's" (l 6) déjà cité: dans ce cas, la hauteur du départ du mouvement sur la syllabe tonique serait perçue comme significative, plus que l'ampleur de la chute ou de la montée. Par contre, pour "lay/the table" (l 1/2) et "Arcopal/tableware" (l 3), le "high fall" semble bien être interprété comme une fin de phrase.

b) Déterminer les accents: (annexe II)

On peut classer les erreurs observées dans quatre catégories. Le segment souligné à tort porte:
- sur une syllabe non-accentuée d'un mot lexical mis en valeur dans le texte (l 1: Sidney, supposed; l 2 drinking; l 9: simpler, Selfridges). Ou plus rarement sur une partie non-accentuée dans un composé comme le dernier mot du texte: Oxford Street (l 9).
- sur une syllabe non-accentuée d'un mot lexical non mis en valeur dans le texte (l 1: supposed, l 9 simpler).
- sur un mot suivant immédiatement une pause syntaxique (l 1 you're, l 3 well, l 6 yet).
- sur un groupe de mots dont l'un seulement peut être accentué (l 1/2 lay the table, l 4 it's tough, l 6 of lovely).

Les sujets testés ont en général souligné plus de syllabes accentuées qu'il n'y en avait en réalité selon le groupe témoin d'anglophones. Pour obtenir des résultats mieux exploitables, il serait nécessaire de reformuler l'énoncé du test en parlant de "syllabe précise" et non plus de "mot". La rapidité du test (une seule écoute sans pause) est certainement aussi à l'origine d'une certaine imprécision dans les schémas soulignés.

Il est difficile de savoir s'il faut attribuer certaines erreurs à une délimitation approximative due à la rapidité de l'exercice, ou bien à une erreur de perception.

En voici deux exemples. Dans le premier cas, l'accent secondaire a pris la place de l'accent principal et la "bonne" réponse était donc: ˈeconomical. C'est ce qu'ont d'ailleurs trouvé pratiquement les deux-tiers des sujets (tableau I); mais comment interpréter les 7 réponses donnant la syllabe 2 ou 4 comme accentuée, alors que les syllabes non-accentuées en anglais sont très souvent réduites ? C'est là une des limites de ce type de test.

Tableau I:

economical	**economical**	economical / economical
	22	7
economical	pas de réponse	TOTAL
2	7	38

Dans le second cas, tough est considéré comme accentué et and est aussi mis en valeur (ce que seulement 40% des sujets ont indiqué), donc la "bonne" réponse était en principe celle où le segment tough and était souligné. C'est du moins ainsi que l'ont perçu les anglophones du groupe témoin. Il existe peut-être un écart entre la perception "linguistique" des anglophones concernés, et la perception plus "phonétique" ou plutôt "acoustique" des sujets francophones qui se seraient plus fiés à leur oreille qu'aux indices syntaxiques. C'est une hypothèse qui permettrait d'expliquer la dispersion des réponses pour ce segment, alors que par ailleurs les résultats étaient relativement bons.

Tableau 2:

tough and pretty	tough and pretty	tough and pretty
6	5	9

tough and pretty	tough and pretty	pas de réponse	TOTAL
4	10	4	38

3) Perspectives:

L'analyse des erreurs amène à se demander s'il existe une seule "bonne" solution, et si l'on doit se fier uniquement à l'intuition de locuteurs natifs. Une étude expérimentale de la fréquence, de l'intensité et de la durée pourra résoudre certaines ambiguïtés [Diana, CELDA, à paraître]; cependant, on peut aussi considérer que si certains indices ne sont pas pris en compte par les locuteurs natifs, c'est qu'ils ne sont pas nécessaires à une bonne compréhension du message, la redondance étant suffisante pour l'assurer.

IV) CONCLUSION:

Le pourcentage de réussite est assez satisfaisant pour la plupart des tests, ce qui peut nous amener à penser que la perception des éléments prosodiques en L2 n'est peut-être pas très différente de celle de la L1. Pour tester cette hypothèse, il conviendrait de mettre au point de nouvelles expériences associant des tests de perception à la fois en L1 et en L2.

D'autre part, il serait utile d'établir des corrélations individuelles entre les tests de perception non

linguistiques et les résultats obtenus pour la parole. Par exemple, les sujets ayant commis le plus d'erreurs au "test d'écoute" sont-ils les mêmes à avoir obtenu de mauvais résultats au test de délimitation des accents ?

Si tel est le cas, peut-on postuler l'existence de plusieurs catégories de sujets ? Quels critères adopter pour les départager [MacCarthy, 1977] ? D'après Landercy et Pankowski [1983] pour ce qui est des sujets musiciens, (p 25): "L'entraînement de l'oreille à la perception de la hauteur augmente considérablement la performance dans la zone fréquentielle où cet apprentissage s'est effectué (80 à 5.000 Hz)". Ces résultats sont-ils valables pour la parole ?

Plus généralement, est-ce que la capacité à percevoir correctement les éléments prosodiques d'une L2 permet de mieux comprendre cette L2 ?

Remerciements:

Je remercie l'équipe de L'Institut de la Communication Parlée de Grenoble (Université Stendhal), et en particulier Véronique Aubergé et Louis-Jean Boé, pour leur aide précieuse lors de la préparation de cet article.

RÉFÉRENCES:

de BOT, C. [1980]
"Intonation teaching and pitch control."*Proceedings of the Institute of Phonetics, Nijmegen* 4, pp 53-63.
BUYSSCHAERT, J. [1985]
"Musical ability and intonation learning."
Phonetics Laboratory, University of Reading. Work in Progress 5, pp 83-95.
DIANA, A. [à paraître]
"Enquête sur la perception de l'anglais oral: compréhension et tests de perception." Actes du 6ème Colloque de Villetaneuse sur l'anglais oral: CELDA, Université de Paris-Nord.
't HART, J. [1981]
"Differential sensitivity to pitch distance, particularly in speech. *Journal of the Acoustical Society of Americaa* 69, pp 811-821.
LANDERCY, A. et PANKOWSKI, D. [1983]
"Perception différentielle de la hauteur." *Revue de Phonétique Appliquée* 65, pp 13-27.
MACCARTHY (1977)
"Auditory and articulatory training for the language teacher and learner." ELT 30, 3, pp 212-219.
O'CONNOR, J.D. and ARNOLD, G.F. [1973]
Intonation of colloquial English (second edition). London: Longman.
RAPHAEL, A. [1971]
"Perception et production des tons et schémas intonatifs de l'anglais dans le premier cycle de l'enseignement supérieur." *Etudes de Linguistique Appliquée* n°3, pp 64-88. Paris: Didier.
RIETVELD, A. C. and GUSSENHOVEN, C. [1985]
"On the relation between pitch excursion size and prominence." *Journal of Phonetics* 13, pp 299-308.
SAUER, H. [1975]
"Discriminationsvermögen im primarschulalter."
Praxis des Neusprachlichen Unterrichts 22, pp 237-245.
SCOTT, Th. [1985]
"Le paysage anglais". In *Mélodie et intonation*, pp 114-118. Besançon: *Bulletin d'Audiophonologie de l'Université de Franche-Comté.*

ANNEXE I:

1) TEST D'ÉCOUTE.

A) Vous allez entendre 2 sons, indiquez si le second est: + haut (+), + bas (-), identique (=):

1)	110-115 Hz	+ 1/2 ton
2)	200-195 Hz	- 1/4 ton
3)	120-115 Hz	- 1/2 ton
4)	200-200 Hz	=
5)	115-110 Hz	- 1/2 ton
6)	195-200 Hz	+ 1/4 ton

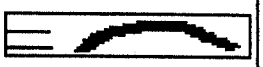
B) Vous allez entendre 3 sons, indiquez lequel est le plus bas:

1)	**110 Hz** / 120 Hz / 115 Hz
2)	195 Hz / **190 Hz** / 200 Hz
3)	120 Hz / **110 Hz** / 115 Hz
4)	**190 Hz** / 200 Hz / 195 Hz

C) Vous allez entendre un son continu, indiquez: s'il monte (+), descend (-) ou reste stable (=).

1)	100-105 Hz	+ 1/2 ton
2)	200-195 Hz	- 1/4 ton
3)	110-104 Hz	- 1/2 ton
4)	150-113 Hz	- 2 tons 1/2
5)	110-105 Hz	- 1/2 ton
6)	200-205 Hz	+ 1/4 ton

D) Dessinez le plus précisément possible le contour des sons suivants:

1)	170 Hz	
(- 4 tons 1/2)	100 Hz	
(+ 1 ton)	112 Hz	
2)	180 Hz	
(+ 1 ton)	200 Hz	
(- 1 ton)	180 Hz	

3)	150 Hz	
(+ 1/2 ton)	160 Hz	
(- 3 tons)	110 Hz	
4)	100 Hz	
(+ 1/2 ton)	105 Hz	
(+ 1 octave)	212 Hz	

2) RÉSULTATS DU TEST:

A) Vous allez entendre 2 sons, indiquez si le second est: + haut, + bas ou identique:

	+	-	=	% bon
1)	**41**	4	3	85%
2)	5	**38**	5	79%
3)	4	**37**	7	77%
4)	1	0	**47**	98%
5)	4	**42**	2	87%
6)	**38**	1	9	79%

B) Vous allez entendre 3 sons, indiquez lequel est le plus bas:

	1	2	3	Rien	1+3	% bon
1)	**29**	7	12			60%
2)	7	**39**	2			81%
3)	2	**36**	9	1	0	75%
4)	**23**	2	12	1	10	48%

C) Vous allez entendre un son continu, indiquez s'il monte (+), descend (-) ou reste stable (=):

	+	-	=	Rien	% bon
1)	**41**	2	5		85%
2)	3	**38**	6	1	79%
3)	4	**42**	2		87%
4)	7	**35**	3	3	73%
5)	1	**42**	4	1	87%
6)	**35**	2	10	1	73%

D) Dessinez le plus précisément possible le contour mélodique des sons suivants:

Évaluation du pourcentage de reproduction exacte des courbes pour 48 réponses:

	(+ 80%)	(de 40 à 60%)	(de 0 à 20%)
1)	3	16	28
2)	17	10	21
3)	6	15	27
4)	21	19	8

ANNEXE II: Test de délimitation.

1) Vous allez entendre le texte ci-dessous, présenté sans ponctuation, lignes numérotées de 1 à 9:
a) Délimitez les phrases du texte:
b) Soulignez les mots ou syllabes accentués:

1. Sidney you're supposed to be helping Mummy lay
2. the table cups are for drinking from not for putting
3. on your head well good job it's Arcopal tableware
4. it's tough and pretty heat and chip resistant too yes I
5. like the colours Arcopal's nice enough to use for
6. best and comes in all sorts of lovely patterns yet it's
7. economical and tough enough to stand up to knocks
8. no Sidney you can't test it on Daddy Arcopal life
9. made simpler available at Selfridges Oxford Street.

2) Texte ponctué et syllabes accentuées soulignées:

Sidney, you're supposed to be helping Mummy lay the table. Cups are for drinking from, not for putting on your head. Well, good job it's Arcopal tableware; it's tough and pretty; heat and chip resistant too. Yes, I like the colours. Arcopal's nice enough to use for best and comes in all sorts of lovely patterns. Yet it's economical and tough enough to stand up to knocks. No Sidney, you can't test it on Daddy.
Arcopal: life made simpler. Available at Selfridges, Oxford Street.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O4

PRODUCTION

PHARYNGALES ET UVULAIRES ARABES SONT DES APPROXIMANTES: CARACTÉRISATION ACOUSTIQUE

YEOU Mohamed & MAEDA Shinji *

LABORATOIRE DE PHONETIQUE DE PARIS (CNRS URA 1027 ET UNIVERSITE LA SORBONNE NOUVELLE)
* DEPARTEMENT SIGNAL, ECOLE NATIONALE SUPERIEURE DES TELECOMMUNICATIONS (CNRS- URA 820)

Résumé

L'objectif de cet article est de comprendre l'acoustique et l'articulatoire des fricatives uvulaires et pharyngales de l'arabe. Une question à laquelle on essaie de répondre: sont-elles vraiment des fricatives ou des approximantes. L'approche est principalement celui de modélisation. On propose des modèles élémentairesqui sont basés sur des fonctions d'aire dérivées de coupes sagittales observées. Les valeurs de formants calculées à partir de ces modèles correspondent bien à celles qui sont mesurées. Les valeurs élevées de Ac et de Ag (l'aire de constriction et l'aire glottique, respectivement), le débit d'air hors des normes pour les fricatives et la structure formantique apparente sur les spectogrammes (surtout de [ʕ,ħ,ʁ]) montrent que ces consonnes ne devraient pas être considérées comme simples fricatives mais approximantes.

Mots-clés: modélisation acoustique, fricatives uvulaires et pharyngales, approximantes, arabe

INTRODUCTION

La relation entre l'articulation et les caractéristiques acoustiques des consonnes fricatives a été étudiée pour plusieurs langues (Fant, 1960; Heinz & Stevens, 1961; Shadle, 1985 etc.). Bien que la modélisation théorique obtenue soit basée sur des suppositions élémentairesen ce qui concerne les mécanismes de production en question, elle est essentielle pour clarifier les paramètres (ex. longueur de la cavité postérieure, position de la constriction, etc.) qui sont déterminants pour la structure acoustique de ces consonnes.

Un des objectifs de cette étude est de comprendre la relation entre l'articulation et les caractéristiques acoustiques des consonnes uvulaires [ʁ,χ] et pharyngales [ʕ,ħ] de l'arabe. Des modèles élémentairesde la fonction d'aire du conduit vocal pendant la production de ces consonnes sont proposés. A partir de ces modèles, les fonctions de transfert correspondantes à chaque configuration sont calculées et l'affiliation entre cavité et formants examinée. Le but de cette idéalisation est de comprendre la structure acoustique qui caractérise ces consonnes. La modélisation est validée par l'analyse acoustique de la parole réelle.

1. FORMULATION DES MODÈLES ÉLÉMENTAIRES

1.1. LES DONNÉES ARTICULATOIRES

La modélisation des consonnes nécessite une connaissance précise d'aire transversalle à chaque point tout au long du conduit vocal, c'est à dire, la fonction d'aire. Malheureusement, il y a un manque de données cinématographiques qui décrivent les dimensions du conduit vocal pendant la production des consonnes uvulaires et pharyngales de l'arabe.

La thèse de Ghazeli (1977) contient des coupes sagittales des fricatives uvulaires et pharyngales qui peuvent être fiablement exploitées pour l'obtention des fonctions d'aire de ces consonnes. Puisque les données sont limitées au contexte de la voyelle /a/, les modèles proposés sont appropriés seulement pour la production de ces consonnes précedées de la voyelle /a/ (cf. Djeradi et *al.*, 1992, pour des coupes sagittales et des fonctions d'aires pour ces même consonnes mais prononcées soutenues et sans contexte vocalique).

Selon Ghazeli (1977), les consonnes pharyngales [ʕ,ħ] ont la même place d'articulation, étant approximativement 3,5 cm au dessus de la glotte dans la cavité laryngo-pharyngienne au niveau de l'épiglotte. La constriction est, par contre, plus étroite pour la sourde [ħ] que pour la sonore [ʕ] (0,3 cm^2 contre 0,4 cm^2). Le larynx est élevé de 0,7 cm dans les deux cas. Les consonnes uvulaires sont caractérisées par un mouvement postéro-supérieur du dos de la langue vers le voile de palais. La place de constriction est plus avancée pour la sourde [χ] que pour la sonore [ʁ].

1.2. FONCTIONS D'AIRE RÉALISTIQUES

Une fonction d'aire est définie comme l'aire cross-sectionnelle du conduit vocal en fonction de la distance de la glotte. Un moyen pour définir cette distance est de trouver la ligne médiane sur la coupe sagittale. Une méthode utilisée pour trouver la ligne

médiane consiste à employer une grille qui découpe le conduit vocal en un certain nombre de sections dont on calcule le centre de gravité. La ligne médiane est constituée de segments joignant ces centres de gravité. Une autre méthode consiste à insérer manuellement des cercles placés l'un près de l'autre et tangentielles au contour du conduit vocal. Le point-centre de chaque cercle définis un point sur la ligne médiane. La distance sagittale qui correspond à ce point est considérée comme le diamètre du cercle (Maeda, 1972; Miller & Fujimura, 1975; Goldstein, 1980; Högberg, 1993). C'est cette dernière méthode qui est utilisée ici (Figure1).

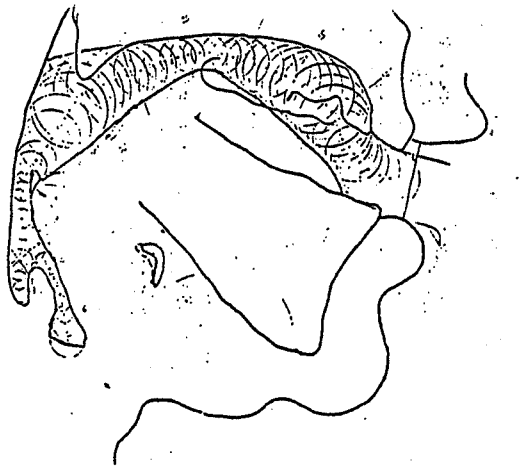


Figure1: cercles inserés à l'intérieur de la coupe sagittale de la consonne pharyngale [ʕ] pour définir la ligne médiane.

Deux profils du conduit vocal, un pour la consonne pharyngale [ʕ] et un autre pour l'uvulaire [ʁ] (Figure2), ont été agrandis 400% pour faciliter l'insertion des cercles. Le diamètre de chaque cercle a été mesuré, ainsi que la longueur des segments joignant les centres de ces cercles. Dans le processus de conversion de la fonction sagittale à la fonction d'aire, une échelle a été définie en fonction du taux d'agrandissement (400 %) et de l'échelle réellement mesurée par Ghazeli (1977). Le passage de la distance sagittale à la fonction d'aire a été fait en appliquant la relation $A=\alpha.d^{\beta}$ où d est la distance sagittale, et α et β sont des coefficients qui diffèrent selon la région du conduit vocal (Heinz & Stevens, 1965). Les fonctions d'aire ont été obtenues par l'application des coefficients déterminés par Fant (1992) et Maeda (1979). Figure 3 montre deux fonctions d'aire dérivées, une pour la consonne uvulaire [ʁ] et une autre pour la consonne pharyngale [ʕ].

Les dimensions de ces fonctions d'aire sont en accord avec les descriptions articulatoires données par Ghazeli pour leurs profils correspondants. Par exemple, la longueur du conduit vocal dans la fonction d'aire pour la consonne pharyngale (16,5 cm) est plus courte que celle pour la consonne uvulaire (17,2 cm) à cause de l'élevation du larynx pendant la production de la consonne pharyngale. La place de constriction pour cette dernière comme est montrée dans la fonction d'aire (Figure3) est de 3,5 cm, en conformité avec ce qui est mesuré par Ghazeli.

Les fonctions de transfert et puis les valeurs de fréquences de formant correspondantes à ces fonctions d'aire ont été calculées par un programme en prenant en compte la radiation aux lèvres, la non-rigidité des parois et le système subglottique (Maeda, 1982). Les fréquences formantiques calculées sont plus au moins proches de celles mesurées pour F1 et F2. Par contre la valeur calculée pour F3 est plus basse. Probablement, ceci est dû à la méthode de passage de la fonction sagittale à la fonction d'aire, qui sous-estime le volume de cavité où il y a changement brusque.

(a)

(b)

Figure2: profils du conduit vocal, d'après Ghazeli (1977), utilisés pour dériver les fonctions d'aire dans Figure3. (a) la pharyngale [ʕ], (b) l'uvulaire [ʁ]

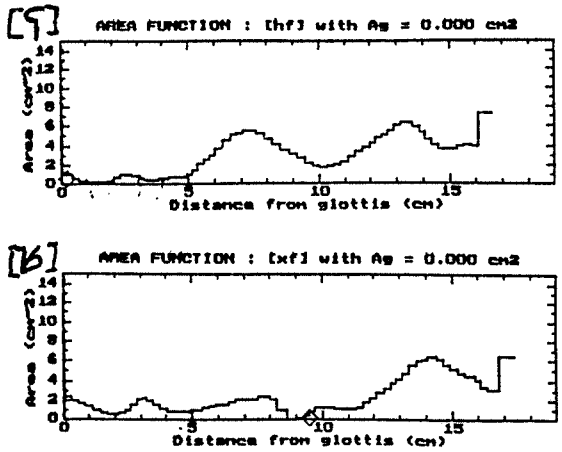


Figure3: fonctions d'aire derivées à partir des profils du conduit vocal de [ʕ] et [ʁ] illustrés dans Figure2.

		calculé	mesuré	
pharyngal [ʕ]			cette étude	Ghazeli (1977)
	F1	674	682-764	850
	F2	1504	1450-1565	1400-1500
	F3	1920	2183-2336	2050-2550
uvulaire [ʁ]				
	F1	472	475-616	500-600
	F2	1216	1137-1252	1200-1300
	F3	1824	2321-2400	2300-2600

Tableau1: comparaison des fréquences mesurées pour [ʕ] et [ʁ] et celles calculées à partir des fonctions d'aire derivées. Les fréquences mesurées sont indiquées par minimum et maximum durant la durée de la consonne.

1.3. FONCTIONS D'AIRES ÉLÉMENTAIRES

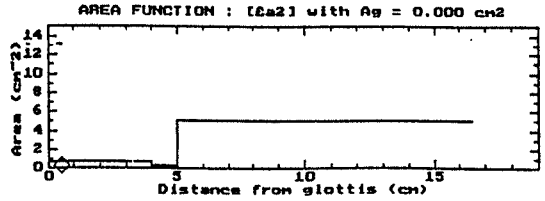
Des modèles élémentairespour la production des consonnes pharyngales et uvulaires sont proposés, en se basant sur les dimensions des fonctions d'aires dérivées comme décrit dans la section 1.2. Le modèle consiste de trois tuyaux uniformes correspondants à la cavité postérieure, la constriction et la cavité antérieure. Les dimensions de ces modèles sont données dans Tableau 2:

	pharyngale	uvulaire
lt (cm)	16,5	17
lc (cm)	1	1
lb (cm)	4	8,5/9,5
lf (cm)	11,5	7,5/6,5
Ac (cm^2)	0,30/0,35	0,20/0,35
Ab (cm^2)	0,7	2
Af (cm^2)	5	5

Tableau 2: lt= longueur totale du tube, lc= longueur de constriction, lb= longueur de la cavité postérieure, lf= longueur de la cavité antérieure, Ac= l'aire de constriction, Ab= l'aire de la cavité postérieure, Af= l'aire de la cavité antérieure.

Ces dimensions sont illustrées par Figure 4. La figure montre que le modèle pharyngal diffère du modèle uvulaire sur plusieurs aspects. Le lieu d'articulation est plus postérieur, l'aire de la cavité postérieure plus étroite et la longueur du tube est plus courte. Pour les consonnes uvulaires, deux modèles sont proposés, un pour l'uvulaire sourde et un autre pour l'uvulaire sonore (la place de constriction pour la sourde est plus avancée que pour la sonore: 9,5 et 8,5 cm de la glotte, respectivement).

(a)



(b)

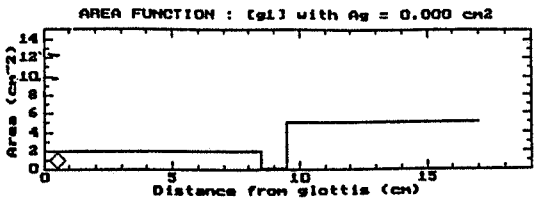


Figure4: modèles élémentairesde fonctions d'aire pour la production des consonnes pharyngales dans (a) et uvulaires dans (b).

2. RÉSULTATS

2.1. FONCTIONS DE TRANSFERT ET FREQUENCES DE FORMANTS

Plusieurs fonctions de transfert, qui correspondent aux fonctions d'aire élementaires, ont été calculées. Les paramètres variables sur lesquels nous jouons sont principalement l'aire à la glotte (Ag) et l'aire de constriction (Ac). La première variable distingue les consonnes sonores (Ag= 0 cm^2) des consonnes sourdes (Ag= 0,20-0,25 cm^2). Le but est d'obtenir une bonne correspondance des fréquences de formant mesurées avec celles obtenues par simulation.

	ʕ	ħ	ʁ	χ
F1 (Hz)	689	784	512/570	------
F2 (Hz)	1493	2040	1160/1206	1328
F3 (Hz)	2181	2648	2218/2304	------
F4 (Hz)	3525	3496	3120/3157	
Ac (cm^2)	0,35	0,30	0,20/0,35	0,20
Ag (cm^2)	0	0,25	0	0,20
Vu (cm^3/s)	-----	596	-----	439

Tableau 3: valeurs des formants et débit d'air (Vu) calculés à partir des fonctions d'aires élémentaires(Figure 4-5) avec l'aire à la constriction (Ac) et l'aire glottique (Ag).

Le tableau montre que F1, F2 et F4 pour les pharyngales [ħ,ʕ] sont plus élevés que pour les uvulaires [ʁ,χ] et que F3 est également plus bas. En plus, les fréquences formantiques des pharyngales sont plus élevées quand la glotte est ouverte que quand elle est fermée. Ceci est dû à l'effet de l'impédance glottique.

Deux valeurs de Ac sont données pour [ʁ] car acoustiquement, on distingue souvent deux parties sur le spectrogramme de [ʁ]: une partie constituée de formants bien définis et une autre de formants très affaiblis et de bruit de friction. Donc, théoriquement pendant la production de la première partie, l'aire de constriction (Ac) est plus large (et Ag≈ 0). On différencie aussi entre les consonnes sourdes et les consonnes sonores au niveau de l'aire de constriction (Ac), qui est plus étroite pour les sourdes que pour les sonores (comme est montré dans Tableau 3).

Les valeurs de Ag et Ac proposées pour les pharyngales et uvulaires sont plus élevées que celles

connues pour les autres fricatives (ex. [s], [ʃ], cf. Stevens, 1971). En effet, il y a plusieurs indices qui montrent qu'on ne peut classer ces consonnes (surtout [ħ,ʕ,ʁ]) comme simples fricatives. Premièrement, une structure formantique est bien évidente sur les spectogrammes correspondant à ces consonnes (voir Figure6). Deuxièment, les valeurs de débit d'air mesurées des consonnes sont soit au dessus ou au dessous de la norme pour les fricatives normales: [ħ]: 595-733 cm^3/s, [χ]: 483 cm^3/s, [ʕ]: 125-165 cm^3/s (Ghazeli, 1977; Butcher & Ahmad, 1987). (cf.125-165 cm^3/s pour les fricatives sonores et 200-500 cm^3/s pour les fricatives sourdes, Stevens, 1971). Ces valeurs mesurées sont utilisées pour la comparaison avec ceux calculées par le programme à partir des modèles élémentaires(cf. Tableau 3). On trouve que les valeurs du débit d'air calculées à partir des modèles sont en bon accord avec celles qui sont mesurées.

.2.2 AFFILIATIONS, EXCITATIONS DES FORMANTS ET LA PLACE DE LA SOURCE.

A partir des modèles élementaires, l'affiliation cavité-formant est la suivante (cf. Stevens & Klatt, 1969):

	cavités	
Formant	pharyngale	uvulaire
F1	antérieure	Helmoltz
F2	Helmoltz	antérieure
F3	antérieure	postérieure
F4	antérieure	antérieure

Dans les fonctions de transfert calculées ci-dessus en haut, on observe que seuls les formants associés avec la cavité antérieure sont excités par la source de bruit quand la glotte est ouverte: (F1, F3 et F4 pour les pharyngales; F2 et F4 pour les uvulaires). En plus, le F2 de la pharyngale [ħ] qui est dû à la resonance de Helmoltz (impliquant la cavité postérieure) se montre aussi excité, un fait qui n'est pas attendu.

Ceci peut être expliqué par le fait que la source de bruit ne se trouve pas à la place de constriction (contrairement au cas des autres consonnes), mais à la glotte. Cette hypothèse apparaît plausible car Ac est plus élevée que Ag (0,30 cm^2 contre 0,25 cm^2). En plus, sur le spectrogramme, on observe une structure formantique bien définie. En effet, la figure (6b) montre que tous les formants (de F1 à F4) sont visibles. Ceci rejoint donc l'hypothèse mentionnée ci-dessus: les consonnes pharyngales et uvulaires, notament [ħ,ʕ,ʁ], ne devraient pas être catégorisées comme simples fricatives, mais comme "approximantes".

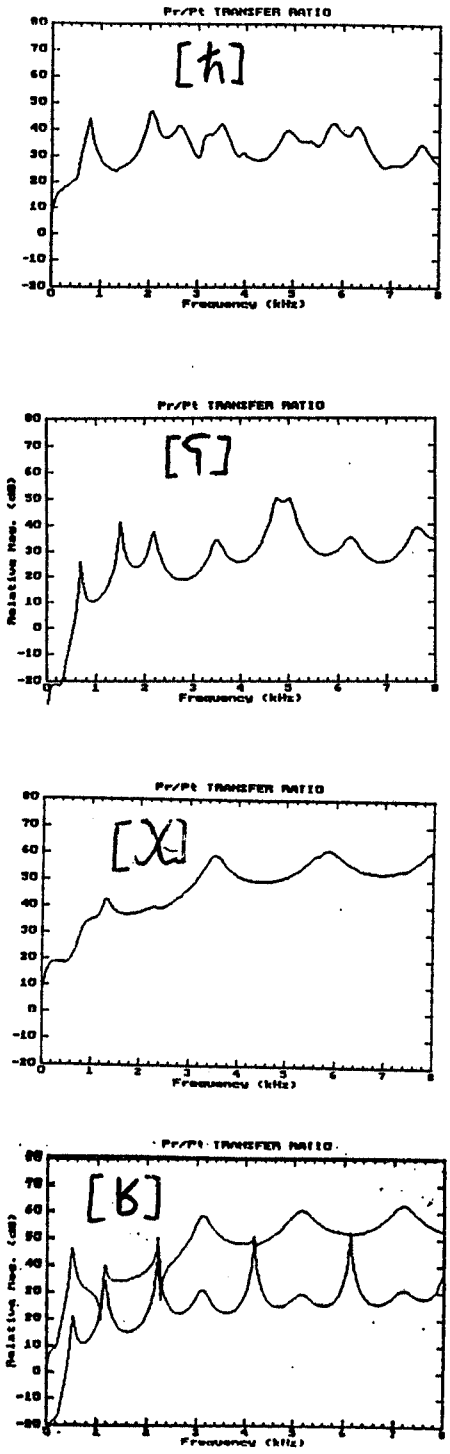


Figure5: fonctions de transfert calculées à partir des fonctions d'aire élementaires. La place de source se trouve pour [ʁ] à la glotte et à 1 cm après la constriction (donc deux fonctions de transfert différentes). Pour [ʕ] et [ħ], la source est à la glotte seulement et pour [χ] elle est à 1 cm après la constriction. En plus, le système subglottique est pris en compte pour [ħ] avec un Ag =0,25 cm^2. Quant à [χ, ʁ, ʕ], Ag = 0,20 cm^2, 0 cm^2, 0 cm^2, respectivement).

3. L'ANALYSE ACOUSTIQUE

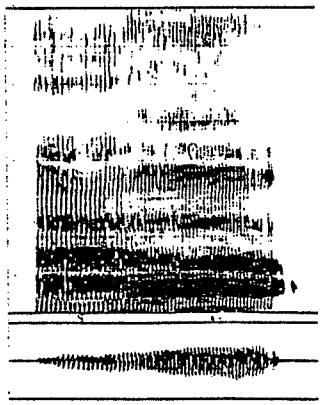
Une analyse acoustique a été faite en parallèle avec la modélisation. Des logatomes de type /CV/ ont été utilisés, où /C/ est [ʁ,χ,ʕ,ħ] et /V/ la voyelle longue [a:]. Ces logatomes ont été inserés dans des phrases cadres de type: "*qa:la CV marratajni*" (il a dit CV deux fois). Quatre locuteurs masculins Marocains ont participé à l'enregistrement. Chaque locuteur a prononcé les phrases cinq fois. Le corpus a fait l'objet d'une analyse acoustique sur PC à l'aide du programme UNICE. Nous avons mesuré les valeurs formantiques au centre des consonnes [ʁ,χ,ʕ,ħ]. Tableau 4 donne les valeurs moyennes des quatre locuteurs.

	ʕ	ħ	ʁ	χ
F1 (Hz)	710	777	616/475	------
F2 (Hz)	1494	1978	1252/1137	1389
F3 (Hz)	2255	2536	2321/2400	------
F4 (Hz)	?	3597	?	?

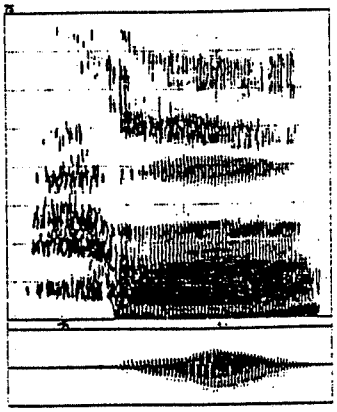
Tableau 4: les valeurs moyennes (sur 4 locuteurs x 5 répétitions) des formants des consonnes pharyngales et uvulaires. Pour [ʁ], on distingue les valeurs de deux parties, une partie au début de la consonne et une à la fin.

Si on compare ces valeurs mesurées avec celles qui sont calculées à partir des fonctions d'aire élémentaires(voir Tableau 3 et 4), on remarque qu'il y une bonne correspondance entre les deux. On a aussi vérifié pour les sourdes quels sont les formants qui sont excités et ceux qui ne le sont pas. pour [χ], on a trouvé que seuls les formants associés à la cavité antérieure sont excités: F2 et F4 (voir Figure 6d). Pour [ħ], tous les formants sont excités (voir Figure 6b). Tous ceci, donc, est en accord avec les prédictions de la modélisation. Une autre prédiction aussi confirmée (comme le montre le tableau 4): Les fréquences formantiques sont augmentées pour la sourde pharyngale [ħ] en comparaison avec la sonore [ʕ].

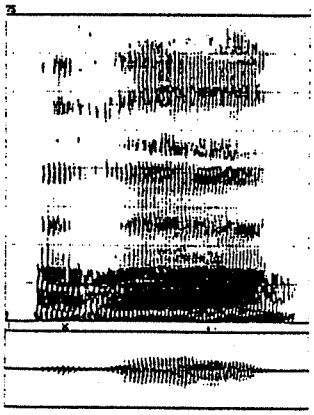
(a)



(b)



(c)



(d)

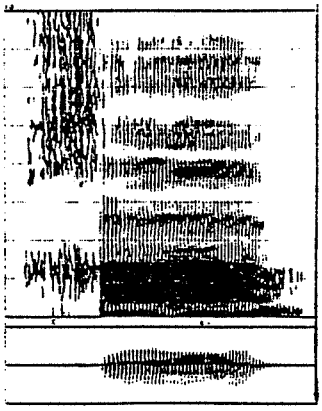


Figure6: spectogrammes correspondants aux consonnes pharyngales et uvulaires suivies de la voyelle [a:]. (a): [ʕa:], (b) [ħa:], (c): [ʁa:] et (d): [χa:].

CONCLUSION

Dans cette étude on a proposé des modèles élémentaires qui sont basés sur des fonctions d'aire dérivées de coupes sagittales du conduit vocal observées. Les valeurs formantiques calculées à partir de ces modèles correspondent bien à celles qui sont mesurées.

La modélisation nous a permis de connaître quels sont les paramètres qui sont acoustiquement importants pour la production et la distinction des consonnes pharyngales et uvulaires. Notament, la position de la constriction, l'aire de la constriction et de la cavité postérieure, et la place de la source. Les valeurs élevées de Ac et de Ag pour modéliser ces consonnes, le débit d'air hors des normes pour les fricatives et la structure formantique apparente sur les spectogrammes (surtout de [ʕ,ħ,ʁ]) montrent que ces consonnes ne devraient pas être considérées comme simples fricatives mais approximantes.

BIBLIOGRAPHIE

Butcher, A. & Ahmad, K. (1987) "Some acoustic and aerodynamic characteristics of pharyngeal consonants in Iraqi arabic," *Phonetica* 44: 156-172.

Djeradi, A., Badin, P. & Guerin, B. (1992) "Effets de couplage subglottique: mesure et modélisation dans le domaine fréquentiel pour les fricatives d'arrière de l'arabe," 19ème *JEP*, Bruxelles.

Fant, G. (1960) *Acoustic Theory of Speech Production.* The Hague, Mouton.

Fant, G. (1992) "Vocal tract area functions of Swedish vowels and a new three-parameter model," *ICSLP 92.*

Ghazeli, S. (1977) *Back Consonants and Backing Coarticulation in Arabic.* Mouton. Ph.D. Thesis, University of Texas.

Goldstein, U.G. (1980) *An Articulatory Model for the Vocal Tracts of Growing Children.* Thèse de Doctorat, Dpt. of Electronic. Eng. and Computer Science, MIT.

Heinz, J.M. & Stevens, K.N. (1961) "On the Properties of voiceless fricative consonants," *J. Acoust. Soc. Am.*, 33: 589-96.

Heinz, J.M. & Stevens, K.N. (1965) "On the relation between lateral cineradiographs, area functions and acoustic spectra of speech," *5th International Congress of Acoustics*, paper A44.

Högberg, J. (1993) "From sagittal distance to area function and male to female scaling of the vocal tract," Rapport de projet. ENST, Paris.

Maeda, S. (1972) "On the conversion of vocal tract X-ray data into formant frequencies," Murray Hill, Bell Laboratories.

Maeda, S. (1979) "An articulatory model of the tongue based on a satistical analysis," *J. Acoust. Soc. Am.*, 65: S22.

Maeda, S. (1982) "A digital simulation method of the vocal tract system," *Speech Communication* 1: 199-229.

Miller, I. & Fujimura, O. (1975) "From tongue model data to sound," *J. Acoust. Soc. Am.*, 57, Suppl. 1.S3.

Shadle, C.H. (1985) *The acoustics of Fricative Consonants.* Ph.D. Thesis, MIT, released as R.L.E. Technical Report 506.

Stevens, K. & Klatt, D. (1969) "Pharyngeal consonants," *Research Laboratory of Electronics MIT Quarterly Proigress Report* 93: 208-216.

Stevens, K. (1971) "Airflow annd turbulence noise for fricative and stop consonants: static considerations," *J. Acoust. Soc. Am.*, 50: 1180-92.

QUELLES STRATÉGIES DE COMPENSATION POUR UNE PERTURBATION LABIALE ?

C. SAVARIAUX *, P. PERRIER * & J.P. ORLIAGUET **

* INSTITUT DE LA COMMUNICATION PARLÉE, URA CNRS 368,
INPG & UNIVERSITÉ STENDHAL. 46 AVENUE FÉLIX VIALLET,
38031 GRENOBLE CEDEX 1, FRANCE.
E-mail : savariaux@icp.grenet.fr, perrier@icp.grenet.fr

** LABORATOIRE DE PSYCHOLOGIE EXPÉRIMENTALE, URA CNRS 665,
UNIVERSITÉ PIERRE MENDES FRANCE. 1491 RUE DES RÉSIDENCES,
38400 S^{T} MARTIN D'HÈRES, FRANCE.

Résumé

Nous présentons ici une expérience pour laquelle les lèvres ont été perturbées, par un tube labial creux de 25 mm de diamètre, lors de la production de la voyelle [u]. Neufs locuteurs naifs ont été étudiés au cours de 3 sessions successives : la première sans perturbation ; la seconde avec le tube labial et la troisième avec le tube labial et en fournissant aux locuteurs des informations sur la stratégie à adopter pour compenser la perturbation. Les analyses sont faites sur la base des rayons X et du patron formantique obtenus. Les compensations réalisées sont évaluées à partir des deux premiers formants. Les résultats obtenus montrent qu'après la seconde session, seulement un locuteur a pu compenser la perturbation. Les autres ayant un second formant très élevé. Dans la troisième session, le comportement de chaque locuteur a été analysé en fonction de l'information donnée : acoustique, articulatoire ou les deux. Les résultats sont analysés en regard des théories classiques de production de la parole.

Mots-clés : Production de la parole, contrôle, perturbation, compensation.

1. INTRODUCTION

Les différentes études acoustiques menées sur la production des voyelles ont montré que plusieurs configurations du conduit vocal, même très différentes, peuvent produire des patrons formantiques quasi identiques (Mermelstein, 1967, Schroeder, 1967, Atal *et al.*, 1978). Cependant, les mesures sur le conduit vocal (Wood, 1979) ou l'exploitation de modèles articulatoires (Boë *et al.*, 1992) tendent à montrer que toutes ces possibilités ne sont pas utilisées dans la parole. Ainsi, pour des réalisations acoustiques de voyelles, émises autant que possible dans des conditions identiques, la *localisation* et la *taille* des constrictions produites avec la langue et/ou les lèvres — donnant la *position* dans l'espace articulatoire de ces voyelles — restent pratiquement inchangées. Pour comprendre ce phénomène, différentes conceptions théoriques du contrôle de la parole ont été proposées :

— l'espace acoustique est l'espace de la tâche (Stevens & Blumstein, 1978) ; mais selon une variante de cette Théorie dite "quantique", les contraintes articulatoires peuvent limiter le nombre de configurations possibles pour un patron formantique donné (Perkell *et al.*, 1993) ;

— l'espace acoustique est bien l'espace d'origine du contrôle de la tâche, mais durant l'apprentissage de la parole des principes d'optimisation, comme par exemple la minimisation d'un coût gestuel, ont entraîné un encodage spécifique de la tâche au niveau géométrique, plus spécifiquement au niveau des constrictions (Gay *et al.*, 1981) ;

— la parole, comme l'a proposé Stetson (1928) serait "*a set of movements made audible*", et l'espace articulatoire est donc l'espace même de la tâche, spécifié directement en "variables du conduit vocal", qui sont aussi des constrictions (Saltzman & Munhall, 1989) ; c'est une conception héritée de la Théorie de l'Action, qui est venue, aux Laboratoires Haskins, en renfort de la Théorie motrice de la perception.

Nous pensons que les expériences de perturbation peuvent apporter des éléments décisifs pour décider entre ces trois théories. Si les perturbations en ligne sont particulièrement bien adaptées à ces gestes transitoires que sont typiquement les consonnes (depuis Abbs & Gracco, 1983, les perturbations les plus nombreuses ont été réalisées sur ce type de segment), le paradigme qui semble le mieux convenir pour tester la résistance aux perturbations des *positions* des voyelles est celui du "pipe speech" (autrement dit "la parole pipe au bec"). Dans cette veine, les premiers résultats intéressants sur la flexibilité des productions vocaliques sont ceux de l'expérience de Lindblom et collègues, dite des *bite-blocks* (Lindblom *et al.*, 1979, Gay *et al.*, 1981) : les locuteurs, malgré l'introduction d'un bloc inter-dentaire (*bite-block*) qui entraîne une position anormale de la mandibule, peuvent aisément compenser la perturbation

et produire les patrons formantiques adéquats, et ceci dès les toutes premières périodes de l'excitation de leur conduit vocal, pour des apertures mandibulaires auxquelles ils n'ont pas eu le temps de s'habituer. Autrement dit, ils parviennent à compenser instantanément, sans apprentissage préalable, en préservant les caractéristiques géométriques des constrictions qui ne sont pas significativement modifiées par rapport aux conditions naturelles. Ces résultats attestent donc du rôle important joué par les constrictions dans la production des voyelles.

Mais cette expérience des *bite-blocks* n'est pas suffisante pour plaider en faveur de l'une ou l'autre des théories citées précédemment. En effet, si les bite-blocks perturbent réellement la position de la *porteuse* des articulateurs qu'est la mandibule, ils n'empêchent pas le locuteur d'atteindre, avec les *effecteurs finaux* que sont la langue et les lèvres, les configurations géométriques du conduit vocal qui sont nécessaires, en l'occurrence les constrictions requises pour produire acoustiquement ces voyelles. Riordan (1977) a été l'une des pionnières de la perturbation à la constriction, en démontrant que des sujets, dont les lèvres (maintenues par un écarteur) étaient empêchées de réaliser la protrusion nécessaire pour produire une voyelle [y], tentaient de compenser, en abaissant leur larynx pour obtenir un allongement *acoustiquement équivalent* du conduit vocal. C'est dans cet esprit que notre expérience a été élaborée pour répondre spécifiquement — non pas à la question du contrôle de la longueur du conduit vocal — mais à celle du contrôle de la *localisation* et de la *taille* de la constriction interne (linguale), qui est la constriction majeure que l'on utilise pour caractériser universellement la *position* des voyelles.

2. LA PERTURBATION PAR TUBE LABIAL

2.1. La nature de la perturbation.

Pour cette étude, le type de perturbation utilisé affecte donc directement les caractéristiques géométriques du conduit vocal à la constriction. La constriction vocalique labiale étant évidemment plus aisée à perturber que celles qui peuvent se situer à l'intérieur du conduit vocal, nous avons choisi de perturber le geste vocalique d'arrondissement d'une voyelle [u]. La perturbation est produite par un tube creux en plexiglass, cylindrique, de 25 mm de diamètre, inséré entre les lèvres du locuteur : le tube labial. La longueur du tube labial est différente pour chaque locuteur, étant choisie de manière à ne pas modifier la longueur totale de son conduit vocal. De plus, un petit *bite-block* (de taille adaptée au comportement du locuteur) a été utilisé de façon à bloquer la mâchoire dans une position naturelle pour cette voyelle, empêchant ainsi toute perturbation produite par un éventuel couplage (mécanique et/ou fonctionnel) entre la mandibule et les lèvres.

2.2. Simulations des conséquences acoustiques et articulatoires de la perturbation.

Les conséquences acoustiques de cette perturbation ont, d'une part, été analysées à l'aide des nomogrammes obtenus par le modèle à 4 tubes de Fant, qui donnent l'évolution des cinq premiers formants lorsque la constriction est déplacée de la glotte jusqu'aux lèvres (Fant, 1960, Badin *et al.*, 1990). Partant de la configuration standard de la voyelle [u] (constriction située environ à 10 cm de la glotte, avec une aire à la constriction environ égale à 0,64 cm^2 et une aire aux lèvres d'environ 0,16 cm^2), nous avons étudié les modifications induites par l'augmentation de l'aire labiale jusqu'à 4 cm^2 (ce qui correspond approximativement à l'aire du tube labial). La principale conséquence acoustique de cette perturbation consiste en une augmentation du second formant F2. Suivant le principe des nomogrammes et en supposant que la longueur du conduit vocal reste constante, il apparaît que le seul moyen de compenser une telle variation du formant F2 est de reculer la constriction à environ 7 cm de la glotte. Ce recul correspond à un changement de localisation de la constriction qui passe nettement d'une zone palato-vélaire à une zone vélo-pharyngale.

D'autre part, les simulations de Bailly *et al.* (1991) et Bailly & Laboissière (1993), utilisant une technique d'inversion acoustique de modèles articulatoires à 5 ou 7 degrés de liberté, ont également montré que la seule possibilité pour diminuer les valeurs du patron formantique créé par une telle ouverture labiale était de reculer la langue.

Ces différentes simulations nous montrent donc qu'il est théoriquement possible de compenser cette perturbation et que cela ne peut se faire que par un recul très prononcé de la langue. Partant de ces prédictions computationnelles, il nous restait à connaître les réactions de locuteurs naïfs à ce type de perturbation.

2.3. Procédure expérimentale.

Neuf locuteurs français (8 hommes et une femme) ont accepté d'être exposés aux rayons X pour les besoins de notre expérience. Cette expérience a été divisée en deux phases successives correspondant à deux conditions différentes. Dans la première phase (condition normale) les locuteurs devaient produire la voyelle [u], avec le *bite-block*, mais sans la perturbation labiale ; ceci nous permettant d'avoir une forme de référence de la voyelle pour chaque locuteur. Dans la seconde phase (condition perturbée) les locuteurs devaient s'efforcer de produire la voyelle [u] avec une qualité auditive aussi bonne que possible et cela malgré le *bite-block* et le tube labial. Pour s'adapter à cette nouvelle situation, 20 essais ont été accordés par locuteur. Pour limiter l'exposition des sujets aux rayons X, seul le dernier essai, que nous avons supposé être le meilleur, a été filmé.

Deux types de données ont ainsi pu être analysées : d'une part les radiographies du conduit vocal, enregistrées dans le plan sagittal ; d'autre part le signal acoustique correspondant, émis pendant au moins une seconde par le locuteur. Pour chaque radio nous avons tracé manuellement le contour du conduit vocal, de la glotte jusqu'aux lèvres ; l'avant du crâne a également été dessiné pour faciliter le repérage anatomique des parties fixes. Pour étudier les différences articulatoires

produites par chaque locuteur, nous avons superposé (en prenant comme référence la partie frontale du crâne et l'incisive supérieure) les tracés obtenus en condition perturbée et en condition normale. D'autre part, une analyse LPC a été réalisée sur le signal acoustique. Ont été retenues les valeurs des trois premiers formants, calculées par moyennage des valeurs prises sur la partie la plus stable du signal. La réussite du résultat acoustique obtenu a été jugée sur la réalisation du patron formantique F1-F2, en fonction des seuils perceptifs donnés par Mermelstein (1978).

2.4. Résultats.

Les différentes réactions observées sont très variables suivant les locuteurs. Elles peuvent néanmoins être catégorisées selon deux classes majeures et trois cas.

• *Sans recul de la langue.* Pour deux locuteurs (cas 1), la forme de la langue reste sensiblement constante (ainsi que le montre la figure 1 pour le locuteur YP ; comme le prédit la théorie acoustique, la valeur du formant F2 est ici trop élevée, cf. tableau 1). Il semble donc que ces deux locuteurs, n'ayant pas trouvé une stratégie de compensation, aient conservé leur configuration interne habituelle du conduit vocal, en dépit de la dégradation du résultat acoustique.

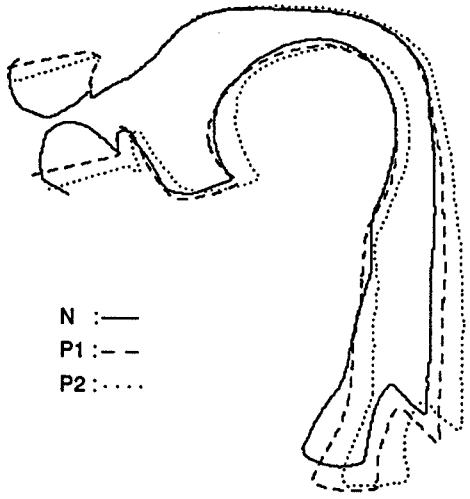


Figure 1 : Tracés du conduit vocal en condition normale (trait plein), perturbée (trait tireté) et perturbée guidée (trait pointillé). Locuteur YP.

• *Avec recul de la langue.* Dans cette classe, deux cas différents peuvent être considérés selon qu'il y a ou non une influence effective du geste articulatoire de compensation sur le patron formantique.

- Pour six de nos locuteurs (cas 2), le geste de recul de la langue (illustré par le locuteur MP sur la figure 2) ne provoque pas de véritable modification de la zone de constriction qui reste palato-vélaire. En conséquence le formant F2 décroît, mais l'amplitude du geste de recul n'est pas suffisante pour atteindre le patron formantique adéquat

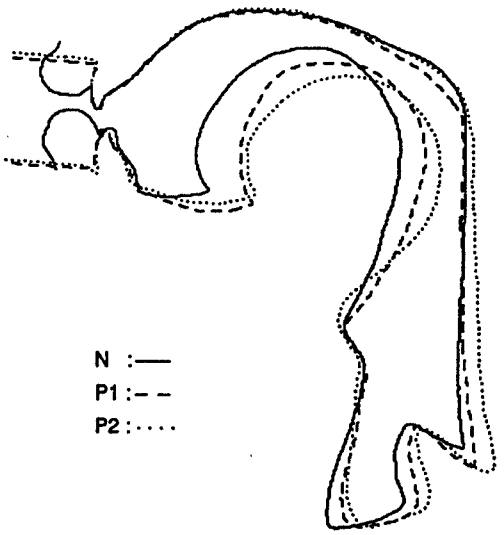


Figure 2 : Tracés du conduit vocal en condition normale (trait plein), perturbée (trait tireté) et perturbée guidée (trait pointillé). Locuteur MP.

- Pour un seul de nos locuteurs (cas 3), le geste de recul de la langue a été suffisant pour faire basculer la constriction de la zone palato-vélaire à la zone vélo-pharyngale (OD, figure 3). En conséquence, comme le prédit d'ailleurs la théorie acoustique, le patron formantique des deux premiers formants correspond bien à celui produit en condition normale.

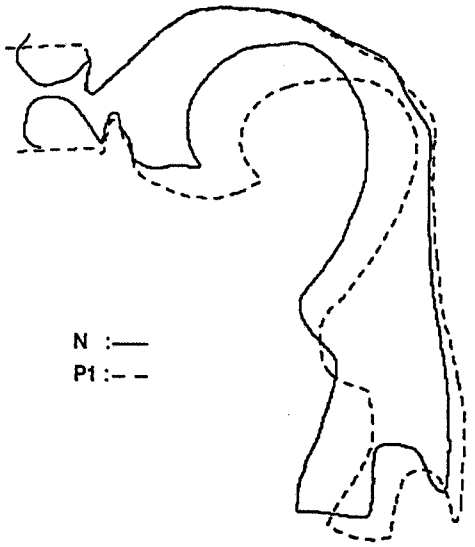


Figure 3 : Tracés du conduit vocal en condition normale (trait plein) et perturbée (trait tireté). Locuteur OD.

En résumé, sept de nos neuf locuteurs ont eu tendance à reculer la langue de manière à compenser l'ouverture labiale. Mais seul l'un d'entre eux a pu produire un geste suffisant pour atteindre le patron formantique adéquat. La majorité des locuteurs semblent donc bien adopter la bonne stratégie, au moins

à l'état de tendance. Mais le fait même qu'un seul locuteur réussisse à produire une réelle compensation acoustique avec cette stratégie — ajouté au fait que deux de nos locuteurs n'ont pratiquement pas modifié leur zone de constriction interne — laisse penser qu'il y a une association forte entre la zone pour produire une voyelle donnée et son patron acoustique. Pour évaluer la force de cette association, nous avons proposé une autre expérience à nos huit locuteurs qui n'avaient pas atteint leur but acoustique lors de la première expérience.

3. L'APPRENTISSAGE AVEC GUIDAGE

3.1 Procédure expérimentale.

Dans cette nouvelle expérience, les locuteurs ont également eu à produire la voyelle [u] en condition perturbée (avec le *bite-block* et le tube labial), mais ils avaient au préalable reçu des informations sur la stratégie à suivre pour réussir à compenser la perturbation. Pour cela, nous avons séparé les locuteurs, sur la base des résultats obtenus lors de la première expérience, en trois groupes différents. Dans le premier groupe appelé *groupe acoustique*, les locuteurs ont écouté le dernier [u] prononcé à la fin de la session précédente et ils ont reçu comme consigne verbale de produire un son plus "grave" que le précédent. Dans le second groupe appelé *groupe articulatoire*, les locuteurs ont visualisé les tracés sagittaux obtenus lors de la première expérience et ils ont reçu comme consigne verbale de déplacer leur langue dans la partie vélo-pharyngale du conduit vocal. Dans le troisième groupe appelé *groupe acoustique et articulatoire* les locuteurs ont reçu les deux types d'information.

De façon à donner le maximum d'informations aux locuteurs du cas 1 (ceux qui n'ont pratiquement pas fait montre d'une capacité de compensation) nous avons placé ces locuteurs dans le groupe *acoustique et articulatoire*. Pour étudier les conséquences de chacun des types d'informations (*acoustique* ou *articulatoire*), nous avons partagé les locuteurs restant entre les deux autres groupes.

Les résultats ont été analysés par rapport aux différentes informations données et ont été comparés à ceux obtenus précédemment.

3.2. Résumé des résultats.

Pour tous les locuteurs du groupe *acoustique*, aucune modification du patron formantique n'a été notée (cf. tableau 1), ce qui correspond à peu ou pas de modification de la position de la constriction (ainsi le locuteur GA — cf. le trait pointillé sur la figure 4 — n'a pas amélioré le recul de sa langue par rapport à la manoeuvre réalisée en condition perturbée sans guidage, cf. le trait tireté).

En revanche, pour tous les locuteurs du groupe *articulatoire*, nous avons noté une diminution plus ou moins importante du formant F2. Cela est la conséquence d'une augmentation du geste de recul de la langue. Néanmoins, cette augmentation reste encore trop faible pour changer suffisamment la zone de constriction et obtenir le patron formantique adéquat.

Les résultats du groupe *acoustique et articulatoire* montrent une faible modification de la position de la constriction (ainsi pour le locuteur YP, figure 1, trait pointillé), mais sans modification du patron formantique.

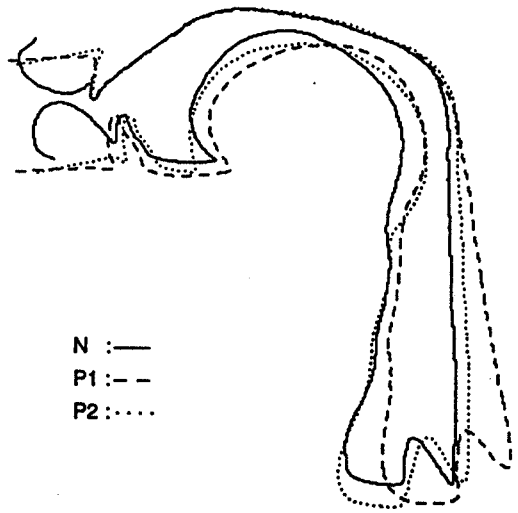


Figure 4 : Tracés du conduit vocal en condition normale (trait plein), perturbée (trait tireté) et perturbée guidée (trait pointillé). Locuteur GA.

4. CONCLUSIONS ET PERSPECTIVES

Pour un système à degrés de liberté en excès comme le conduit vocal, la question classique de l'équivalence motrice, se double d'une autre question essentielle pour la communication parlée, celle de l'*équivalence acoustique* — entendue au sens premier de la psychophysique, c'est-à-dire sans faire intervenir dans un premier temps les variations permises par le système phonémique. Au contraire des perturbations mandibulaires — qui ont été les plus couronnées de succès pour les voyelles (Lindblom *et al.*, 1979, Gay *et al.*, 1981), mais qui se sont révélées, selon nous, être des perturbations uniquement motrices —, la perturbation que nous avons provoquée sur la constriction aux lèvres d'une voyelle arrondie-type, comme [u], par un tube labial de 25 mm de diamètre, a directement des conséquences acoustiques très importantes. Pourtant, la théorie acoustique prédit que la compensation est possible, à condition de reculer de manière appropriée le lieu de la constriction linguale.

Effectivement l'un de nos neuf sujets français naïfs a été capable de compenser d'emblée, selon la stratégie prédite, cette forte perturbation du patron acoustique du [u], produite par une ouverture excessive aux lèvres. Six autres ont amorcé une stratégie dans le même sens ; mais celle-ci n'a pas provoqué un effet acoustique suffisant. Enfin les deux locuteurs restants n'ont esquissé aucun comportement notable de compensation.

Ainsi, contrairement à une simple perturbation motrice sur un articulateur distal pour la production des voyelles, comme l'est la mandibule, la capacité de compenser une perturbation sur des effecteurs finaux, en l'occurrence les lèvres, semble être bien partagée par la majorité des locuteurs, mais ne serait pas habile au point d'être toujours suivie de succès. Elle est même difficile à apprendre pour la plupart des sujets, tout spécialement lorsque les informations de guidage sont délivrées en termes auditifs. Il est important de noter que parmi tous les essais qui ont échoué acoustiquement, nous n'avons pas relevé un seul comportement "aberrant", c'est-à-dire une réaction qui se serait trompée de direction du point de vue des effets acoustiques : ainsi aucun de nos sujets n'a avancé la langue.

A la suite de ces expériences, la connaissance procédurale implicite des sujets parlants, en ce qui concerne la production des voyelles, nous semble pouvoir se caractériser ainsi : la majorité d'entre eux laissent penser qu'ils ont connaissance de la stratégie nécessaire à la réussite de la compensation, aucun locuteur, de surcroît, ne démontre qu'il choisit une démarche contraire. Ceci plaide en faveur de l'hypothèse d'une connaissance implicite de la relation entre articulatoire et acoustique, le *forward model* proposé par Jordan & Rumelhart, 1992 (cf. aussi pour la parole le *modèle direct interne* de Laboissière, 1992). Les différences comportementales entre les locuteurs traduiraient ainsi la capacité de chacun à rendre explicite cette connaissance au niveau articulatoire, c'est-à-dire à passer du «*que faut-il faire ?*» au «*comment faire ?*».

Ces expériences permettent-elles de décider quelle est la théorie la plus plausible pour le contrôle des voyelles dans la parole ? La capacité de compensation, même limitée, de la majorité des sujets ne favorise pas une interprétation du contrôle en termes uniquement articulatoires (Saltzman & Munhall, 1989). Pas plus que les possibilités de guidage articulatoire ne conviennent à un contrôle en termes uniquement acoustiques (des contraintes proprement articulatoires pourraient par contre bien jouer pour expliquer la difficulté de réorganisation de la tâche, Perkell *et al.*, 1993). La thèse qui parait le plus plausible à la suite de nos expériences pourrait être la suivante :

(1) la tâche vocalique est définie dans l'espace acoustique en termes de valeurs formantiques ;
(2) l'apprentissage de la parole conduit les locuteurs à associer une forme du conduit vocal à un patron acoustique donné (notion déjà proposée par Gay *et al.*, 1981 : "*The target of a vowel is coded neurophysiologically in terms of area-function related information ...*") ;
(3) en cas de perturbation majeure (incapacité pour les locuteurs à faire appel à cette association), l'habileté de chaque sujet à réagir, dépend de sa capacité à réexploiter une relation articulatori-acoustique déjà éprouvée, et ainsi mémorisée dans son modèle direct interne.

Locuteurs		Conditions	Formants		
			F1	F2	F3
	OD	N	327 (2)	769 (20)	1840 (9)
		P1	334 (2)	734 (29)	2212 (7)
	BC	N	303 (13)	759 (27)	1881 (26)
		P1	330 (5)	927 (11)	2118 (19)
		P2	334 (2)	963 (21)	2237 (15)
Groupe	CH	N	224 (3)	632 (12)	1994 (34)
		P1	364 (3)	842 (14)	2279 (7)
Acoustique		P2	356 (2)	833 (17)	2176 (35)
	GA	N	236 (9)	720 (5)	1630 (69)
		P1	206 (3)	945 (9)	2007 (62)
		P2	314 (24)	1022 (34)	1992 (52)
	JY	N	262 (7)	716 (17)	2055 (315)
		P1	362 (7)	1259 (8)	2552 (12)
		P2	331 (1)	1021 (8)	2868 (73)
Groupe	LJ	N	327 (4)	677 (13)	2150 (17)
		P1	418 (11)	945 (7)	2334 (30)
Articulatoire		P2	366 (14)	804 (6)	2396 (18)
	MP	N	272 (20)	686 (24)	2003 (42)
		P1	284 (4)	853 (5)	2272 (10)
		P2	292 (5)	785 (18)	2330 (151)
Groupe	ML	N	288 (8)	663 (7)	1683 (63)
		P1	388 (17)	1170 (11)	2527 (22)
Acoustique		P2	359 (8)	1079 (4)	2333 (11)
et	YP	N	301 (5)	668 (11)	2194 (18)
		P1	354 (4)	930 (22)	2229 (19)
Articulatoire		P2	304 (6)	1043 (22)	2215 (12)

Tableau 1 : Moyenne en Hz des trois premiers formants (et écart type) pour chaque locuteur, en conditions normale (N), perturbée (P1) et perturbée guidée (P2). Pour la répartition en trois groupes, cf. texte.

REMERCIEMENTS

Nous remercions C. Abry et L.J. Boë pour leurs conseils et les discussions fructueuses que nous avons eues avec eux. Remerciements également à tous les locuteurs qui ont subi ces expériences.

Ces travaux ont été en partie financés par le projet ESPRIT/BR n° 6975, Speech Maps.

RÉFÉRENCES

ABBS J. H. & GRACCO V. L. (1983). Sensorimotor actions in the control of multimovement speech gestures. *Trends in Neuroscience, 6,* 393-395.

ATAL B. S., CHANG J. J., MATHEWS M. V. & TUKEY J. W. (1978). Inversion of articulatory-to-acoustic tranformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am., 63,* 1535-1555.

BADIN P., PERRIER P., BOË L.J., & ABRY C. (1990). Vocalic nomograms: acoustic and articulatory considerations upon formant convergences. *J. Acoust. Soc. Am., 87,* 1290-1300.

BAILLY G., ABRY C., BOË L.J., LABOISSIÈRE R., PERRIER P. & SCHWARTZ J.L. (1992). Inversion and speech recognition. In J. R. Vandewalle, M. Boite, A. Moonen & M. Oosterlinck (Eds.), *Signal processing VI: theories and applications* (Vol.1, pp. 159-164). Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

BAILLY G., JORDAN M., MANTAKAS M., SCHWARTZ J.L., BACH M. & OLESEN M. (1990). Simulation of vocalic gestures using an articulatory model driven by a sequential neural network. *J. Acoust. Soc. Am., 87,* S1, 105.

BAILLY G., LABOISSIÈRE R. & SCHWARTZ J.L. (1991). Formant trajectories as audible gestures: an alternative for speech synthesis. *J. of Phonetics, 19,* 9-23.

BAILLY G. & LABOISSIÈRE R. (1993). Learning coarticulation and compensation for selected vowels to vowels with a control model. In G. Bailly (Ed.), *Speech Maps Report for year 1* (Volume 4). ESPRIT-BR Project n°6975, European Community.

BOË L.J., PERRIER P. & BAILLY G. (1992). The geometric variables of the vocal tract controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion. *J. of Phonetics, 20,* 27-38.

FANT G. (1960). *Acoustic theory of speech production.* The Hague, The Netherlands: Mouton.

GAY T., LINDBLOM B. & LUBKER J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *J. Acoust. Soc. Am., 69,* 802-810.

GAY T., BOË L.J. & PERRIER P. (1992). Acoustic and perceptual effects of changes in vocal tract constrictions for vowels. *J. Acoust. Soc. Am., 92,* 1301-1309.

JORDAN M. I. & RUMELHART D. E. (1992). Forwards models: Supervised learning with a distal teacher. *Cognitive Sciences, 16,* 316-354.

LABOISSIÈRE R. (1992). *Préliminaires pour une robotique de la communication parlée : inversion et contrôle d'un modèle articulatoire du conduit vocal.* Thèse de l'Institut National Polytechnique de Grenoble. Université de Grenoble, France.

LINDBLOM B., LUBKER J. & GAY T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *J. of Phonetics, 7,* 147-161.

MERMELSTEIN P. (1967). Determination of the vocal-tract shape from measured formant frequencies. *J. Acoust. Soc. Am., 41,* 1283-1294.

MERMELSTEIN P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *J. Acoust. Soc. Am., 63,* 572-580.

PERKELL J. S., MATTHIES M. L., SVIRSKY M. A. & JORDAN M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study. *J. Acoust. Soc. Am. ,93,* 2948-2961.

RIORDAN C. (1977). Control of vocal-tract length in speech. *J. Acoust. Soc. Am, 62,* 998-1002.

SALTZMAN E. L. & MUNHALL K. G. (1989). A Dynamical approach to gesture patterning in speech production. *Ecological Psychology, 4,* 1615-1623.

SCHROEDER M. (1967). Determination of the geometry of the vocal tract by acoustic measurements. *J. Acoust. Soc. Am., 41,* 1002-1010.

STETSON R. H. (1928). Motor Phonetics. A study of speech movements in action. *Archives Néerlandaises de phonétique expérimentale, 3,* 1-216 (Nouvelle édition par Kelso J. A. S. & Munhall K. G. (1987). Boston, Massachussets: Little, Brown and Company).

STEVENS K. N. & BLUMSTEIN S. E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am., 64,* 1358-1368.

WOOD S. (1979). A radiographic analysis of constriction locations for vowels. *J. of Phonetics, 7,* 25-43.

La coarticulation mandibulaire comme principe d'organisation syllabique (Et si la mandibule syllabisait ALIS ?)

N. RHARDISSE & C. ABRY

ICP URA CNRS n° 368 INPG/ENSERG UNIVERSITÉ STENDHAL
BP 25 38 040 GRENOBLE CEDEX 9 FRANCE

Résumé

Lindblom (1983) et Keating *et al.* (1990) ont corrélé l'*adhérence* des consonnes au noyau vocalique à leur déplacement mandibulaire et celui-ci à la variabilité *relative* (σ/m), comme indice de *coarticulabilité* du segment. Le suivi des positions mandibulaires verticales de deux sujets français et de deux marocains a été analysé. Les segments-tests choisis pour cet espace sont [i, a] et [s, l]. Nos résultats soutiennent un ordre mandibulaire de hauteur croissante : [a] < [l] ≤ [i] < [s]. Cet ordre semble résister étonnamment à la différence de langue. Nous obtenons, en outre, pour tous nos locuteurs confondus un ordre de *coarticulabilité relative* parfaitement corrélé à cette échelle. Les implications de ces deux échelles — de *hauteur* et de *coarticulabilité relative mandibulaires* — devraient permettre de repenser l'économie de la syllabe.

Mots-clés : *mandibule, coarticulation, syllabe.*

1. INTRODUCTION

Syllabe, hauteur et coarticulation mandibulaires

Dans ses propositions sur l'économie articulatoire de la syllabe, Lindblom (1983) formule une première hypothèse, selon laquelle les propriétés biomécaniques du système mandibulaire peuvent sans doute rendre compte de la régulation syllabique des gestes d'ouverture et de fermeture du tractus vocal autour de 6 Hz, à travers les différentes langues, hypothèse qui reprend les suggestions de Brodda (1979, p. 66) à partir de la durée moyenne des syllabes (Sigurd, 1973), et les résultats des expériences préliminaires de Sorokin *et al.* (1980).

Il avance en outre une seconde hypothèse (cf. déjà Lindblom, 1979), pour tenter d'expliquer la constitution des syllabes complexes à partir des ajustements CV, à savoir que la propension des sons à se grouper en syllabes dépendrait fortement de la hauteur de leur position mandibulaire, qui déterminerait ainsi une «coarticulation "propensity"». Plus précisément, plus la résistance des consonnes à la coarticulation mandibulaire de hauteur vocalique est forte (p. ex. pour [s] dans [sa]), plus celles-ci auraient tendance à occuper les places syllabiques les moins *adhérentes* au noyau vocalique (cf. *strate*). On voit ici l'application du principe selon lequel les syllabes optimiseraient la coarticulation de leurs segments, en tendant à éloigner graduellement les plus différents pour obtenir un décours du geste syllabique lisse de préférence.

Keating (1983) a remis en question les résultats de Lindblom (1983) sur le suédois, qui ne correspondraient pas à l'affirmation de ce dernier, selon laquelle les consonnes s'accommoderaient plus ou moins à leur noyau vocalique. Elle soutient au contraire que certaines consonnes, comme [s], accommodent les voyelles et les autres consonnes moins exigeantes sur le recrutement de la mandibule. Finalement, Keating, Lindblom, Lubker & Kreiman (1990) ont remis la seconde hypothèse de Lindblom (1983) à l'épreuve comparée de l'anglais et du suédois, en mesurant les déplacements mandibulaires (exprimés en % du max. d'ouverture référencé par rapport à l'occlusion) pour les consonnes [f,b,t,d,s,n,l,r,k,h] réalisées dans des environnements vocaliques [a-a], [e-e] et [i-i]. Les résultats montrent, globalement pour ces deux langues (Fig. 5), que les segments les plus fermés sont les plus variables et les segments les plus ouverts sont les plus stables, à condition que la variabilité soit exprimée *relativement*, c'est-à-dire par le coefficient de variation (écart-type/moyenne) : c'est ce que nous avons rebaptisé *coarticulabilité relative*. Il y a donc une corrélation entre les deux échelles (r = 0,80 sur les *rangs*, Keating *et al.*, 1990, p. 89) : celle d'ouverture ou de hauteur mandibulaire et celle de coarticulabilité ou de résistance relative à la coarticulation. Ce résultat soutient donc la position de Lindblom, comme quoi ce sont plutôt les consonnes qui accommodent leur hauteur mandibulaire à celle des voyelles environnantes.

Hauteur et variabilité des consonnes

Pourtant, on remarquera que, globalement (anglais et suédois confondus), l'échelle de hauteur mandibulaire ne sépare pas nettement les consonnes des voyelles (cf. Fig. 5). Ainsi la voyelle la plus haute [i] se trouve être plus élevée que [l], [k] et [h]. Pour ces segments, la coarticulabilité relative (σ/m) de [k] est plus proche de

celle de [a] et de [i], alors que celle de [l] l'est plutôt des consonnes comme [t] ou [f] ; celle de [h] se situant entre [l] et [k]. (On notera que cette consonne la plus basse [h], est celle qui est précisément réputée ne pas avoir de position supraglottique définie). Il est donc de première importance de connaître sur ces échelles les relations de proximité du noyau vocalique et des marges consonantiques. En conséquence, notre corpus comprendra les valeurs réputées extrêmes de hauteur mandibulaire [s] *vs.* [a] et les valeurs extrêmes de hauteur mandibulaire vocalique [i] *vs.* [a]. Quelle est donc la consonne la plus proche de la voyelle la plus haute ? [k], [l] et [b] sont les meilleurs candidats chez Keating *et al.* (1990). Mais il se trouve que [l] est finalement la plus proche en pourcentage d'ouverture ([l] = 39,9 , [b] = 34,2, [k] = 40,1 et [i] = 37,8).

L'intérêt de rechercher la position mandibulaire de la consonne [l] nous semble en fait résider dans l'ensemble des résultats suivants. Dans une étude antérieure sur l'arabe marocain (Rhardisse & Sock, 1991), nous avons pu mettre en évidence que la position mandibulaire des [l] simples, et surtout des [ll] géminés, se trouvait (et ceci surtout dans la vitesse d'élocution la plus lente) être extrêmement proche de la position de la voyelle [i].

Mais est-il vraiment possible d'attribuer au phonème [l] une cible mandibulaire? Lors d'une étude réalisée sur le [l] américain, Varney (1933) soulignait la variabilité de ce phonème, non seulement du point de vue de la mandibule (Simon 1967, Bothorel 1978), mais aussi d'autres articulateurs tels que la langue, les lèvres et le larynx. Sensible à l'entourage phonique, le [l] varie selon le sujet qui le produit et selon la langue dans laquelle il est réalisé (Straka, 1965). Chafcouloff (1980), dans une étude acoustique du phonème français, avance que la latérale n'est contrainte qu'à l'apex, tandis que les autres parties de la langue restent libres de prendre toutes les configurations selon la voyelle contiguë. Dans une analyse acoustique portant sur les deux allophones de l'anglais *Received Pronounciation*, [l] clair "palatalisé" et [l] sombre "vélarisé", Bladon & Al-Bamerni (1976 ; Bladon, 1979) remarquent que ce dernier résiste davantage à la coarticulation.

Les expériences menées sur les fricatives définissent le [s] comme étant moins variable que la latérale, quoiqu'il ne soit pas insensible à son entourage. Ainsi Straka (1965) démontre que le [s] peut se labialiser et se sonoriser au contact d'éléments labiaux et voisés. Mais si le [s] subit l'influence des voyelles labiales, sa position mandibulaire n'est altérée par aucune articulation en contact (Bothorel, 1978). Ainsi Bladon & Nolan (1977) attestent qu'en anglais la fricative [s] est bien très résistante à la coarticulation.

Nous allons, à partir de ces différents points de vue, nous interroger sur les variations des cibles mandibulaires (revue in Bognar, 1983), pour ces latérales et ces fricatives de l'arabe marocain et du français (pour une comparaison sur ces dernières, cf., entre autres, Delattre *et al.*, 1989, et Jomaa & Abry, 1993, pour l'arabe).

2. MÉTHODE

Pour mettre en évidence la hauteur mandibulaire du [l] et du [s] au contact des positions haute [i] et basse [a], et tester leur résistance mutuelle à la coarticulation, nous avons choisi de faire varier deux facteurs influant sur le paramètre temps : l'un métrique, joue sur la quantité consonantique ou gémination ; l'autre jouera sur la vitesse d'élocution (cf. *infra*).

Nous avons en conséquence établi un corpus, tiré de chacune des deux langues (arabe et français), qui comporte huit paires minimales. Les oppositions choisies en arabe sont les suivantes : [ala] «unité de mesure», [alla] «il a fait mijoter», [ali] «il a fait frire», [alli] «fais frire», [ila] «si», [illa] «sauf», [ili] «médisance» et [illi] «logatome». [asa] «il a enduré», [assa] (logatome), [asi] «méchant», [assi], [isa], [issa], [isi] et [issi] sont des logatomes. Ces différents items ont été insérés dans la phrase porteuse [a:l — lamin] «Lamin a dit». En français, les items sont précédés par [al] et se présentent ainsi : [ala] «Al hâla», [alla] «Al hâl(e) là», [ali] «Al a Li», [alli] «Al hâl(e) Li», [ila] «Al il a», [illa] «Al il l'a», [ili] «Al y lit», [illi] «Al il lit» ; [asa] «Al a ça», [assa] «Al a c(e) sas», [asi] «Al a Si», [assi] «Al a c(e) Si», [isa] «Al hissa», [issa] «Al hiss(e) ça» [isi] «Al y scie», [issi] «Al hiss(e) Si». Al, Li et Si sont des noms propres. Toutes les réalisations sont en modalité interrogative. Notons qu'en arabe, la première voyelle de chaque item est précédée d'une légère glottalisation. L'opposition de quantité est en l'occurrence partie intégrante du système phonologique de l'arabe marocain, alors qu'en français la gémination consonantique n'est réalisée que par le contact de consonnes hétéromorphémiques.

Enregistrement

Ont servi de sujets, deux locuteurs féminins de l'arabe de Fès, l'auteur N. R. et sa sœur (soit N et A), et deux locuteurs français (C, féminin et F, masculin), tous étudiants de niveau doctoral.

Les items ont été enregistrés, en chambre sourde, en ordre aléatoire, à raison de douze répétitions par item ; une première série avec une consigne de vitesse d'élocution normale (dite conversationnelle) et une seconde série à un débit plus rapide. L'appareillage utilisé se compose d'un kinésiographe mandibulaire (Myotronics K5AR), permettant de suivre la position verticale d'un aimant collé contre les incisives inférieures du sujet et d'un enregistreur FM (Lallouache &Worley, 1988 ; l'excursion de l'aimant est cadrée dans la partie linéaire du kinésiographe, Kies, 1992). Sur les douze répétitions enregistrées, nous n'avons pu en retenir régulièrement que neuf, certains des signaux

articulatoires ou acoustiques présentant différentes défectuosités.

Numérisation, édition et mesures
Nous avons numérisé en stéréo les signaux acoustiques et les signaux articulatoires à 8kHz avec le logiciel Audiomedia (sur Macintosh). Puis nous les avons édités sous Signalyze. Comme le signal de déplacement vertical de la mandibule est de basse fréquence, nous avons réduit son taux d'échantillonnage à 500 Hz. Sur ce signal nous avons mesuré la hauteur mandibulaire de chaque phonème, aux alentours de son centre acoustique, en nous référant au minimum minimorum d'ouverture mandibulaire pour chaque locuteur. Les valeurs numériques obtenues sur Signalyze ont été converties en millimètres (grâce à l'étalonnage effectué par bloc dentaires, à l'enregistrement), nous donnant ainsi 3456 mesures.

3. RÉSULTATS

Nous donnerons ici seulement les résultats d'ensemble, sans rendre compte pour l'instant de l'analyse de la variance pour les principaux effets (de contexte, de gémination et de débit) que permet de tester notre corpus. Nous donnerons ainsi le portrait global de chaque locuteur en nous basant sur les valeurs de moyenne et d'écart-type pour construire ses échelles de rang en hauteur et coarticulabilité (sans regroupement statistique, pour rester comparables à Keating *et al.*, 1990). Puis nous observerons les tendances générales des deux langues, arabe et français, après avoir normalisé les valeurs moyennes et les coefficients de variation des quatre locuteurs. Ces résultats seront commentés au fur et à mesure par rapport à ceux de l'anglais et du suédois.

Hauteur et coarticulabilité : stratégies des locuteurs (Tab. 1 et Figs 1-4)

	a	l	i	s
A	7,17 *0,150*	4,62 *0,227*	3,74 *0,172*	2,50 *0,262*
N	11,96 *0,159*	9,32 *0,122*	7,51 *0,177*	5,05 *0,188*
C	10,33 *0,198*	9,18 *0,136*	6,73 *0,308*	2,45 *0,475*
F	8,95 *0,231*	6,18 *0,251*	5,86 *0,233*	3,56 *0.333*

Tableau 1.— Moyenne d'ouverture mandibulaire (en mm) et coefficient de variation (σ/m) des segments [a, l, i, s], pour les locuteurs arabes (A et N) et français (C et F).

Partant du tableau 1, qui donne les résultats globaux, par locuteur, on notera que l'excursion en ouverture ou déplacement moyen de la mandibule est naturellement variable par individu : elle va de 4,67 mm en moyenne pour le locuteur A, à 7,88 mm pour C.
Mais sur l'échelle de hauteur mandibulaire, nous avons pour tous l'ordre croissant suivant : [a] < [l] < [i] < [s]. Ainsi cet ordre vaut pour l'arabe, le français et le suédois confondu avec l'anglais (Fig. 5). De fait, en moyennant séparément les résultats de Keating *et al.* (1990, p. 82), seul le suédois présente cet ordre (de [a] à [s] en % : 61,4 ; 48,2 ; 38,3 ; 21) ; alors que l'anglais a un [l] vélarisé légèrement plus haut que [i] (respectivement : 62,7 ; 31,6 ; 37,4 ; 15,9). En résumé, un «ordre» [a] < [l] ≤ [i] < [s] reflète la proximité possible de [l] et [i], rencontrée aussi en arabe (pour A plus que N) et en français (seulement pour F).
En ce qui concerne la coarticulabilité relative, nous sommes en présence de deux comportements différents, qui ne sont pas liés à la langue : celui de N (Fig. 2) et C (Fig. 4) ; celui de A (Fig. 1) et F (Fig. 3), qui est lui semblable, en terme de rangs, aux résultats de Keating *et al.* (1990), présentés sur la Fig. 5.
On aura remarqué, pourtant, que la corrélation entre hauteur et coarticulabilité est assez bien retrouvée, les segments ne s'écartant pas de plus d'un rang de la diagonale croissante des graphiques (à titre indicatif, r = 0,80 sur ces quatre segments, Figs 1-4 ; r = 0,69 sur les 13 segments de la Fig. 5, avec r = 0,80 sur les *rangs*, cf. *supra*). [s] est le plus stable de tous dans son extrême coarticulabilité et [l], sans aucun doute, celui qui change le plus de rang ; [i] et [a] étant assez stables de ce point de vue.

Hauteur et coarticulabilité : tendances générales (Tab. 2 et Fig. 6)

	a	l	i	s
A	100 *0,230*	45,44 *0,484*	26,62 *0,526*	0 ∞
N	100 *0,275*	61,92 *0,258*	35,81 *0,531*	0 ∞
C	100 *0,260*	85,41 *0,187*	54,32 *0,479*	0 ∞
F	100 *0,384*	48,63 *0,596*	42,67 *0,586*	0 ∞
m% σ%/m%	100 *0,287*	60,35 *0,381*	39,86 *0,531*	0 ∞

Tableau 2.— Moyenne d'ouverture mandibulaire normalisée (en % du déplacement moy. du loc., cf. Tab. 1) et coefficient de variation (avec σ normalisé en % du même déplacement, soit σ%/m%), avec moyenne de m% et de σ%/m%, tous locuteurs confondus (mêmes segments et locuteurs qu'au Tab. 1).

Pour obtenir le comportement général de ces quatre segments, nous avons normalisé les moyennes et les cœfficients de variation de la hauteur mandibulaire pour chaque locuteur (Tab. 2). La normalisation d'une valeur moyenne est obtenue par la différence de cette valeur

avec la valeur moyenne minimale (qui sert d'origine), divisée par la différence entre la valeur moy. max. et la valeur moy. min., soit le déplacement mandibulaire moyen du locuteur (résultat exprimé en %). La normalisation du coefficient de variation (σ/m) est réalisée en deux temps : en divisant l'écart-type σ par le même déplacement moyen (normalisation exprimée en %) ; puis en divisant ce σ% par la valeur de la hauteur moyenne normalisée m%. (N.B. la convention d'une origine normalisée à 0%, donne évidemment des coefficients qui tendent vers l'infini.)

Les valeurs de la dernière ligne du tableau 2 sont représentées sur la Fig. 6. On y retrouve tous locuteurs et langues confondus une corrélation parfaite entre hauteur et variabilité relative. Nos résultats confirment donc la thèse de Keating *et al.* (1990) selon laquelle les échelles mandibulaires de position et de variabilité peuvent être utilisées comme indices pour comprendre l'économie de la syllabe.

4. DISCUSSION

Cette approche de la question des cibles et de leur variation — envisagée du point de vue de la porteuse mandibulaire, dont la synergie intégrative modèle la cohérence syllabique — par la construction d'indices de hauteur et de coarticulabilité relative, permet de renouveler à la fois la question des cibles et celle de la constituance, sur des principes articulatoires, dans le cycle de base de la parole.

L'échelle de hauteur mandibulaire croissante : [a] < [l] ≤ [i] < [s] résiste étonnamment à la différence de langue, et de locuteur — et pour anticiper sur des résultats que nous n'avons pas présentés ici, de métrique ([l] *vs.* [ll] ; [s] *vs.* [ss]) et de débit.

La corrélation entre les deux échelles de hauteur mandibulaire et de coarticulabilité, qui se retrouve globalement pour des locuteurs de langues aussi différentes que l'anglais, l'arabe, le français et le suédois doit répondre à des principes de contrôle. La fonction de cette relation est ainsi à rapporter au contrôle moteur, mais aussi au niveau du contrôle articulatori-acoustique (Stevens, 1989), sachant en particulier que la sensibilité acoustique dépend du logarithme de l'aire à la constriction (Fant, 1960), une sensibilité à laquelle contribue la mandibule, en synergie, même si elle n'est pas l'effecteur final de la constriction. En outre cette relation permet d'expliquer à la fois la propension des consonnes à être distantes du noyau vocalique, tout en rendant praticable la coarticulation des éléments distanciables lorsqu'ils sont, comme c'est le cas syllabiquement le plus fréquent, en contact direct avec les voyelles.

Reprenant dernièrement la très vénérable, et toujours actuelle, question de l'échelle de sonorité, Clements (1989) propose avec cinq traits, *bas*, *ouvert* (cf. pour ce dernier, Clements, 1993), *vocoïde*, *approximant* et *sonorant*, une échelle à six degrés, allant des voyelles basses [a] aux obstruantes [s, t], en passant par les voyelles moyennes [e], les voyelles hautes [i], les liquides [r, l] (pour les latérales, cf. la proposition d'en faire un trait de classe majeur par McDonough, 1993) et les occlusives nasales [n]. Cette échelle devrait tenir compte du caractère spécifique des fricatives comme [s] parmi les obstruantes. En effet, contrairement à «l'échelle largement consensuelle» (résumée, par exemple, in Klein, 1993, p. 102 sq.), qui donne ces fricatives comme plus sonores que les plosives, la hauteur mandibulaire pourrait fournir un argument en faveur de la «bonne» formation syllabique des groupes du type [spl, str,...], à en juger par la mise en relation que donne Lindblom (1983) de ses mesures mandibulaires avec la hiérarchie de sonorité de Elert (1970), pour le suédois. (Si la phonologie ne continuait pas à négliger la mandibule dans la spécification de ses traits ou même de ses gestes, sur la base d'arguments de variabilité ou de compensation. Nous répondons ici aux premiers ; la réponse aux seconds est la structure coordinative, domaine de la synergie. Notons que même dans la phonologie articulatoire, qui introduit pourtant la synergie, la mandibule n'entre pas dans la partition gestuelle, mais dans la coordination interarticulateurs. D'où l'absence d'arguments internes lorsqu'il s'agit de reprendre les propositions métriques des autres pour noter la structure syllabique dans une telle partition, Browman & Goldstein, 1986.)

Comme nous le proposons sur une base mandibulaire, l'échelle de sonorité peut aussi jouer, en phonologie, entre les voyelles, Clements (1989, p. 19) avançant même que la différence entre [i], [e] et [a] est en quelque sorte parallèle à celle entre [d], [n] et [l]. C'est une position qui se trouve soutenue en l'occurrence par les analyses de Steriade (1984) et Dell & Elmedlaoui (1985), qui montrent, pour le roumain et le berbère, que la distinction entre les voyelles hautes et non hautes peut, elle aussi, être cruciale pour la syllabification.

Il resterait à introduire des principes d'optimisation intra- et inter-syllabiques proposés dans la simulation de Lindblom *et al.* (1984). L'exploration à la Maddieson (1993, Maddieson & Precoda, 1992 ; pour les latérales, cf. Maddieson & Emmorey, 1984) de bases de données importantes devrait nous permettre de retrouver à l'œuvre, dans les syllabes préférentielles des langues, le principe de négociation entre locuteur et auditeur (sinon le rapport bénéfice perceptif/coût articulatoire) pour modeler la syllabe.

Remerciements à Tahar Lallouache, Rabah Kies et Alain Arnal, pour le soin qu'ils ont pris à rendre nos signaux fiables, à Caroline Smith pour ses références bibliographiques, et à Louis-Jean Boë, pour sa permanente disponibilité à tous nos problèmes.

5. RÉFÉRENCES

BLADON, R. (1979), The production of laterals : Some articulatory properties and their acoustic implications. *Current issues in the Phonetic Sciences.* Amsterdam, John Benjamins, 501-508.

BLADON, R. & AL-BAMERNI A. (1976), Coarticulation resistance in English /l/. *J. Phonetics* 4, 137-150.

BLADON, R. & NOLAN, F. (1977), A video-fluorographic investigation of tip and blade alveolars in English. *J. Phonetics* 5, 185-193.

BOGNAR, E. (1983), Directions des recherches phonétiques consacrées à la mâchoire. *Bull. de l'Inst. de Phon. de Grenoble* 12, 1-45.

BOTHOREL, A. (1978), *Étude phonétique et phonologique du breton parlé à Argol (Sud-Finistère).* Thèse d'État, Strasbourg.

BRODDA, B. (1979), Något om de Svenska Ordens Fonotax och Morfotax. *Papers from the Institute of Linguistics, University of Stockholm* 38, 63-108.

BROWMAN, C. & GOLDSTEIN, L. (1986), Towards an articulatory phonology. *Phonology* 3, 219-252.

CHAFCOULOFF, M. (1980), Les caractéristiques acoustiques de [j, ɥ, w, l, r] en français. *Trav. de l'Inst. de Phon. d' Aix* 7, 7-56.

CLEMENTS, G. (1989), A unified set of features for consonants and vowels. ms., Institut de Phonétique, Paris.

CLEMENTS, G. (1993), Un modèle hiérarchique de l'aperture vocalique : le cas Bantou. In B. Laks et M. Plénat (Eds.), *De natura sonorum*, Presses Universitaires de Vincennes, 23-64.

DELL, F. & ELMEDLAOUI, M. (1985), Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *Journal of African Languages and Linguistics* 7, 105-130.

DELATTRE, C. JOMAA, M. AL-DOSSARI, A. & WORLEY, C. (1989), The phasing of the jaw in consonant and vowel lengthening : Arabic and French patterns. *European Conference on Speech Communication and Technology* 2, 416-419.

ELERT, C. (1970), *Ljud och Ord i Svenskan.* Almqvist & Wiksell, Stockholm.

FANT, G. (1960), *Acoustic theory of speech production.* Mouton, The Hague.

JOMAA, M & ABRY, C. (1993), Jaw phasings and velocity profiles in Arabic. *European Conference on Speech Communication and Technology* 1, 545-548.

KEATING, P. (1983), Comments on the jaw and syllable structure. *J. Phonetics* 11, 401-406.

KEATING, P. LINDBLOM, B. LUBKER, J. & KREIMAN, J. (1990), Jaw position in English and Swedish VCVs. *U.C.L.A. W.P.* 74, 77-95.

KIES, R. (1992), Étude des signaux mandibulaires. Plan de phase dans les occlusives. *Rapport Interne* I.C.P., Grenoble.

KLEIN, M. (1993), La syllabe comme interface de la production et de la réception phoniques. In B. Laks et M. Plénat (Eds.), *De natura sonorum*, Presses Universitaires de Vincennes, 101-141.

LALLOUACHE, M. & WORLEY C. (1988), Saisie, édition, et traitement d'images et de signaux articulatoires lèvres et mâchoire. *Journal d'Acoustique* 1, 215-220.

LINDBLOM, B. (1979), Some phonetic null hypotheses for a biological theory of language. *9th I.C.Ph.S.* 2, 33-40.

LINDBLOM, B. (1983), Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech.* Springer Verlag, N.-Y.- Heidelberg, 217-245.

LINDBLOM, B. MacNEILAGE, P. & STUDDERT-KENNEDY, M. (1984), Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie and O. Dahl (Eds.), *Explanation of languages universals.* Mouton, The Hague, 181-203.

MADDIESON, I. (1993), The structure of segment sequences. *U.C.L.A. W.P.* 83, 1-7.

MADDIESON, I. & EMMOREY, K. (1984), Is there a valid distinction between voiceless lateral approximants and fricatives ? *Phonetics* 41, 181-190.

MADDIESON, I. & PRECODA, K. (1992), Phonetic models and syllable structure. *Phonology* 9, 45-60.

McDONOUGH, J. (1993), The phonological representation of laterals. *U.C.L.A. W.P.* 83, 19-32.

RHARDISSE, N. & SOCK, R. (1991), Résistance à la coarticulation mandibulaire du [l] simple et du [ll] double en arabe marocain. *12th I.C.Ph.S.* 2, 50-54.

SIGURD, B. (1973), Maximum rate and minimal duration of repeated syllables. *Language and Speech* 16, (4), 373-395.

SIMON, P. (1967), *Les consonnes françaises. Mouvements et positions articulatoires à la lumière de la radiocinématographie.* Klincksieck, Paris.

SOROKIN, V. GAY, T. & EWAN, W. (1980), Some biomechanical correlates of jaw movement. *J.A.S.A.* 68 (S1), S32 (A).

STERIADE, D. (1984), Glides and vowels in Rumanian. *Proceedings of the 10th Meeting of the Berkeley Linguistics Society*, University of California, Berkeley, 47-64.

STEVENS, K. (1989), On the quantal nature of speech. *J. Phonetics* 17, 3-45.

STRAKA, G. (1965), *Album phonétique.* Presses de l'Université Laval, Québec.

VARNEY, J. (1933), *Études sur l'l américain.* Presses Universitaires de France, Paris.

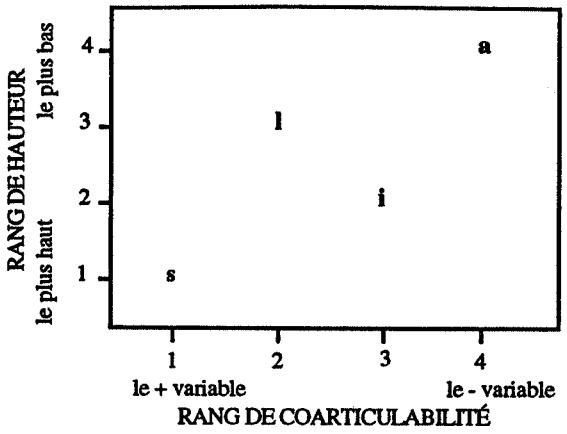


Fig. 1.— Arabe (locuteur A)

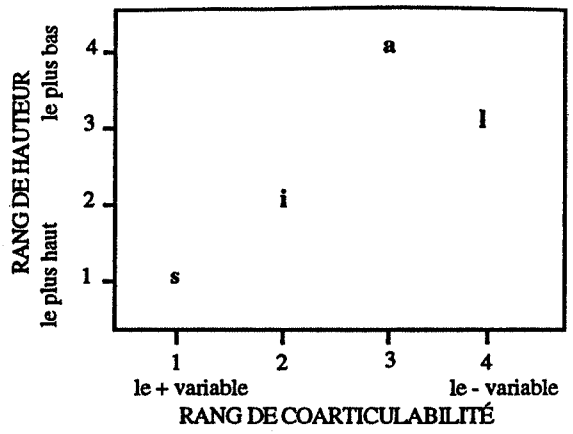


Fig. 2.— Arabe (locuteur N)

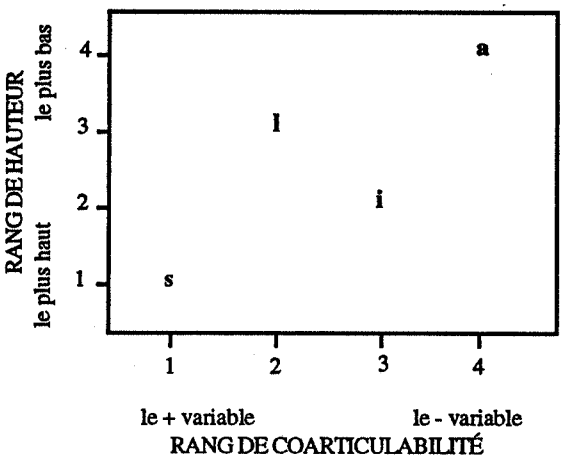


Fig. 3.— Français (locuteur F)

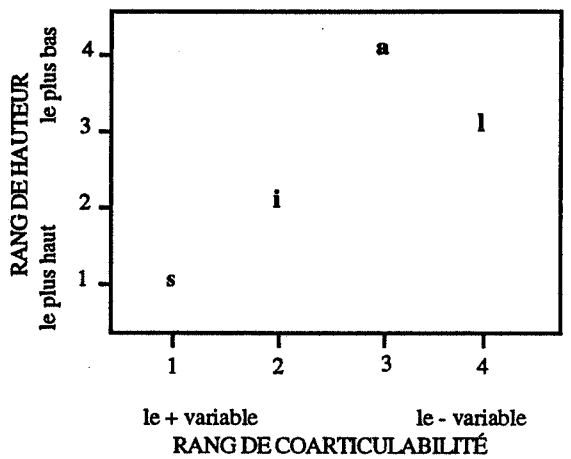


Fig. 4.— Français (locuteur C)

Figs 1-4.— Rang de hauteur mandibulaire (d'après les moyennes m, en mm de déplacement, référencé au min. des min. du locuteur) en fonction du rang de coarticulabilité relative (d'après les coefficients de variation σ/m). Résultats par locuteur.

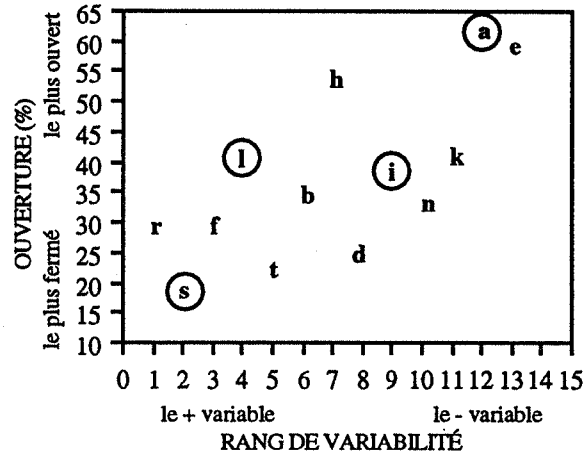


Fig. 5.— Ouverture mandibulaire (m en % du max. de déplacement) en fonction du rang de variabilité (d'après σ/m). Anglais et suédois confondus (adapté de Keating *et al.*, 1990). Nos segments-tests sont encerclés.

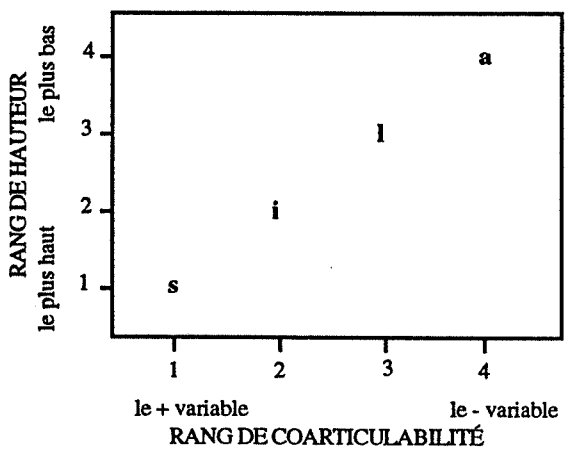


Fig. 6.— Rang de hauteur mandibulaire (d'après m normalisé en % du déplacement moy. du loc.) en fonction du rang de coarticulabilité relative (avec σ normalisé en % du même déplacement, soit σ%/m%). Tous locuteurs arabes et français confondus.

ESTIMATION DE TRAJECTOIRES ARTICULATOIRES A L'AIDE D'UN RESEAU DE NEURONES

Martine GEORGE[1], Paul JOSPA[2] & Alain SOQUET[3]

INSTITUT DES LANGUES VIVANTES ET DE PHONETIQUE, CP 110,
UNIVERSITE LIBRE DE BRUXELLES,
Avenue F.D. Roosevelt, 50, B-1050 Bruxelles, Belgique

Résumé

Cet article propose une méthode de génération de transitions pseudo-articulatoires à partir de trajectoires des trois premiers formants. Un réseau de neurones contrôle les paramètres d'un modèle de la fonction d'aire (paramètres articulatoires) sur base des 3 premiers formants. Le modèle de fonction d'aire choisi est celui à 8 Régions Distinctives. Pour entraîner le réseau, deux ensembles de données sont à notre disposition : le premier contient pour chacune des 11 voyelles du français, les 3 premiers formants et les paramètres articulatoires correspondants; le second est caractérisé par 17 triplets de formants extraits de transitions vocaliques en parole naturelle. Quatre modes d'apprentissage du réseau sont comparés : supervisé, autonome, hybride, supervisé-hybride. Le réseau est à même de générer les trajectoires des paramètres articulatoires en traitant chronologiquement, point par point, chaque triplet de formants.

Mots-clés : *réseau de neurones, formants, trajectoires articulatoires.*

INTRODUCTION

Ces dernières années, les réseaux connexionnistes ont été utilisés dans le cadre de l'inversion acoustico-articulatoire (Bailly, Bach, Laboissière & Olesen, 1990; Laboissière, Schwartz & Bailly, 1991; Shiraï & Kobayashi, 1991; Rahim, Goodyear, Kleijn, Schroeter & Sondhi, 1993). (Saerens & Soquet, 1991; Soquet, Saerens & Jospa, 1990) ont réalisé une inversion acoustico-articulatoire d'un modèle de conduit vocal en combinant la théorie des Régions et des Modes Distinctifs (Mrayati, Carré et Guérin, 1988) avec un réseau de neurones artificiels. L'algorithme utilisé, une variante de l'algorithme de rétropropagation de l'erreur, a été développé dans le cadre de la commande de processus (Saerens, Soquet, Renders & Bersini, 1992; Saerens 1991), et présente de fortes similarités avec des algorithmes de commande adaptative. Il nécessite une connaissance à priori du Jacobien du processus physique (Soquet, Saerens & Jospa, 1990) ou d'une approximation de celui-ci (Jospa, Soquet & Saerens, 1992). Dans le travail présenté ici, nous utilisons une méthode précise et rapide de calcul du Jacobien basée sur une formulation variationnelle, proposée par (Jospa 1994). Cette dernière fournit explicitement, en plus des formants, et avec très peu de calculs supplémentaires, les fonctions de sensibilité et le Jacobien associés à la configuration articulatoire proposée par le réseau à chaque itération. La méthode d'inversion acoustico-articulatoire utilisée est celle développée par (Soquet, Saerens & Jospa, 1990). Celle-ci combine un réseau connexionniste multicouches, un modèle de la fonction d'aire et la méthode variationnelle du calcul des formants et du Jacobien. Le modèle de la fonction d'aire choisi est celui à 8 Régions Distinctives (*Figure 1*) proposé par (Mrayati, Carré & Guérin, 1988).

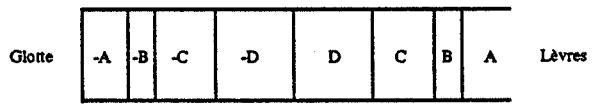


Figure 1 : Tube acoustique subdivisé en 8 régions de longueurs inégales.

Après avoir décrit quatre modes d'apprentissage du réseau et une méthode de génération de transitions articulatoires au départ de transitions formantiques, nous présentons des comparaisons entre les configurations articulatoires obtenues pour ces différents modes dans le cas de deux transitions vocaliques [i a] et [uɔ].

APPRENTISSAGES

Le réseau de neurones utilisé comprend 3 couches. La couche d'entrée reçoit les formants et comprend 3 unités. Nous avons adopté une architecture à une seule couche cachée de 18 unités. Ce choix fournit une marge de manoeuvre suffisante au réseau pour assurer l'apprentissage de toutes les données et un temps de convergence acceptable. La couche de contrôle

[1]Bourse ULB, [2]Science Program, N° SCI*- CT92-0786, [3]Bourse BNB.

comprend 9 unités correspondant aux sections des 8 régions et à la longueur du conduit vocal. Les coefficients d'apprentissage et de moment du réseau ont été fixés respectivement à 0.3 et 0.4.

Nous avons étudié 4 modes d'apprentissage du réseau : autonome (A), supervisé (S), hybride (H), supervisé-hybride (S+H).

Autonome (Saerens, Soquet 1991) : dans la littérature, on trouve couramment des données relatives aux trois premiers formants des 11 voyelles du français (Mrayati, 1976; Majid, Boe & Perrier, 1986). Un premier mode d'apprentissage, a été réalisé sur ces données. Il se déroule selon le cycle d'apprentissage suivant (*Figure 2*):

1. Le réseau reçoit en entrée les formants cibles des voyelles.
2. Les sorties du réseau fournissent les paramètres articulatoires (aires des 8 régions et longueur du conduit).
3. Les formants, fonctions de sensibilité et Jacobiens de la transformation acoustico-articulatoire sont calculés par la méthode variationnelle pour la configuration articulatoire proposée par le réseau.
4. Les différences entre les formants obtenus et les formants cibles sont utilisées pour calculer l'erreur en sortie. Les nouvelles valeurs du jacobien sont alors utilisées pour rétropropager cette erreur jusqu'aux entrées du réseau et pour corriger les poids des connections.

Ces 4 étapes sont répétées jusqu'à ce que l'écart entre les formants générés et les formants cibles soit inférieur à un seuil.

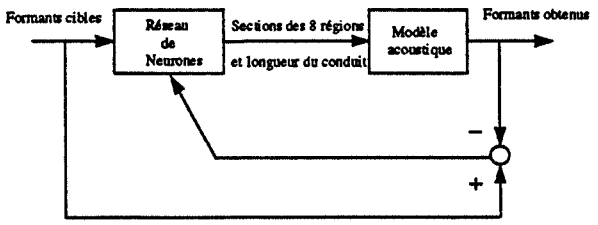


Figure 2 : Apprentissage autonome. Le réseau doit générer les sections des tubes qui permettent de produire les formants cibles.

Cet apprentissage est dit autonome car il ne nécessite au départ que la connaissance des entrées du réseau (les formants). Notons également que le nombre de triplets de formants à faire apprendre au réseau n'est limité que par un souci du temps de convergence. (Soquet, Saerens & Jospa, 1992) ont réalisé cet entraînement sur l'entièreté de l'espace vocalique.

Supervisé : une autre manière d'entraîner le réseau consiste en un apprentissage de type supervisé (Rahim, Goodyear, Kleijn, Schroeter & Sondhi, 1993). En plus des fréquences formantiques, il est nécessaire dans ce cas de connaître a priori les fonctions d'aire qui leur correspondent. De manière à avoir un grand nombre de données d'apprentissage réalistes, il est souhaitable de disposer d'un grand nombre de données articulatoires issues de l'observation directe (par rayons X par exemple) (Bothorel, Simon, Wioland, Zerling, 1986). Il est ensuite possible grâce à une optimisation (au sens de Newton-Raphson) de déduire de ces données les paramètres du modèle articulatoire choisi pour l'inversion; à savoir, dans notre cas, les sections des 8 régions et la longueur totale du conduit vocal. Le processus d'apprentissage supervisé se déroule comme suit :

1. Le réseau reçoit en entrée les formants cibles des voyelles.
2. Les sorties du réseau fournissent les aires et la longueur du tube
3. Les différences entre ces valeurs et les configurations articulatoires cibles sont utilisées pour calculer l'erreur en sortie, qui une fois rétropropagée jusqu'aux entrées du réseau, permet l'ajustement des poids de ce dernier.

Cette méthode permet d'introduire dans le réseau une connaissance explicite de nature articulatoire. L'espoir qui sous-tend cette démarche est de fournir au réseau des informations suffisantes pour lui permettre, lors de la génération des trajectoires, de proposer des configurations articulatoires réalistes. L'inconvénient est la nécessité d'avoir un grand nombre de couples formants-paramètres articulatoires issus de l'observation directe du conduit vocal et permettant de couvrir une large plage de l'espace vocalique.

Hybride : étant donné la difficulté de constituer un ensemble d'apprentissage suffisant pour fonctionner en mode supervisé, il semble intéressant de combiner les deux apprentissages précédents. Ceci nous amène à considérer un troisième mode d'apprentissage. Ce dernier, que nous appelons hybride, se déroule comme suit (*Figure 3*) :

1. En fonction de la nature des données qui lui sont présentées (formants seuls ou formants et paramètres articulatoires), le réseau détermine son mode d'apprentissage (autonome ou supervisé).
2. En autonome, il applique (A); en supervisé, (S).

Notons que les trois modes d'apprentissage précités initialisent les poids du réseau de manière aléatoire.

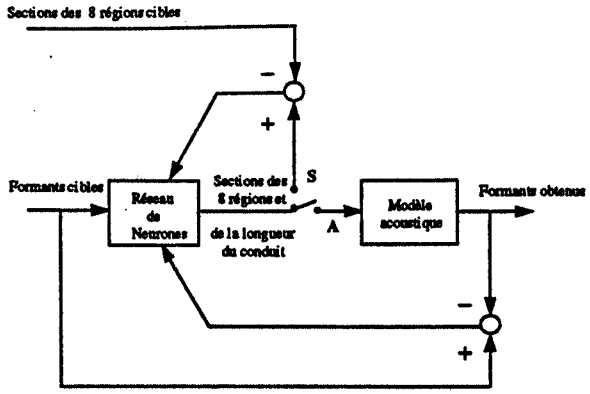


Figure 3 : Apprentissage hybride. Le réseau fonctionnera en autonome (A) ou en supervisé (S) selon le type de données qui lui sont présentées.

Supervisé-hybride : un quatrième mode d'apprentissage, variante du précédent, consiste à entraîner au préalable le réseau en supervisé et à enchaîner avec un apprentissage hybride. Ceci est motivé par le fait que l'apprentissage supervisé fournit une connaissance a priori au réseau qui pourrait être utilisé à meilleur escient par l'apprentissage hybride qu'une initialisation aléatoire.

GENERATION DE TRAJECTOIRES ARTICULATOIRES

La génération des trajectoires articulatoires est réalisée par un réseau issu d'un des 4 modes d'apprentissage décrits ci-dessus. Les trajectoires formantiques de transitions V1V2 sont traitées point par point. Les triplets de formants de chacun des échantillons temporels sont successivement présentés au réseau. Pour chaque triplet de formants, une optimisation (locale) garantit l'adéquation entre les formants cibles et ceux obtenus à la sortie du conduit vocal. Cette optimisation est réalisée d'une manière similaire à l'apprentissage autonome. Ainsi, pour chaque transition formantique, on obtient les trajectoires articulatoires correspondant aux différents paramètres de contrôle du conduit vocal.

RESULTATS ET DISCUSSION

Nous avons réalisé chacun des 4 modes d'apprentissage à trois reprises. Pour l'autonome, les données de départ sont les trois premiers formants des 11 voyelles du Français proposées par (Majid, Boe & Perrier, 1986). Ces mêmes données formantiques complétées par les configurations articulatoires correspondantes constituent l'ensemble de données utilisées pour l'apprentissage supervisé. Pour l'apprentissage hybride et supervisé-hybride, les données sont d'une part, celles de l'apprentissage supervisé et d'autre part, 17 triplets de formants extraits de transitions formantiques de parole naturelle et choisis de manière à constituer des valeurs intermédiaires relativement bien distribuées dans l'espace vocalique (*Figure 4*).

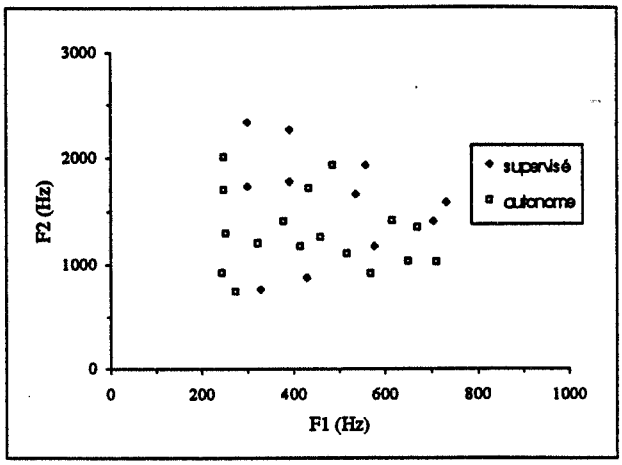


Figure 4 : Distribution dans le plan F1-F2 des donnés formantiques utilisées pour l'apprentissage hybride.

Les seuils de convergence du réseau utilisé en autonome sont les critères absolus suivants :

$$\left|F_1^{cible} - F_1^{obtenu}\right| \leq 30Hz \quad (1)$$

$$\left|F_2^{cible} - F_2^{obtenu}\right| \leq 40Hz \quad (2)$$

$$\left|F_3^{cible} - F_3^{obtenu}\right| \leq 50Hz \quad (3)$$

où les F_i^{cible} sont les formants cibles et les F_i^{obtenu} les formants obtenus (i = 1, 2, 3).

Les critères de convergence du réseau utilisé en supervisé sont définis comme suit:

$$\left|\sqrt{A_i^{cible}} - \sqrt{A_i^{obtenu}}\right| \leq k\sqrt{A_i^{cible}} \quad (4)$$

$$\left|L^{cible} - L^{obtenu}\right| \leq 0.02L^{cible} \quad (5)$$

où les A_i^{cible}, A_i^{obtenu} sont respectivement les fonctions d'aires cibles et obtenues pour chacune des régions; de même L^{cible} et L^{obtenu} sont respectivement les longueurs du conduit cible et du conduit obtenu. Un seuil de convergence: $k = 0.03$ est choisi pour la région de constriction principale ainsi que pour la région labiale; un seuil moins exigeant : $k = 0.06$ est choisi pour les autres régions.

Le *Tableau 1* reprend les nombres moyen de cycles d'apprentissage observés pour chacun des 4 modes. Les temps de convergence des algorithmes d'apprentissage sont relativement différents. L'apprentissage autonome

est de loin le plus rapide et l'apprentissage hybride de loin le plus lent. C'est cette constatation qui nous a amené à effectuer un apprentissage supervisé-hybride. En tenant compte des temps d'apprentissage de la partie supervisée et ensuite de la partie hybride (avec initialisation des poids du réseau avec les valeurs résultant de l'apprentissage supervisé), l'entraînement du réseau en mode supervisé-hybride converge plus rapidement que l'apprentissage hybride.

Modes	Taille des ensembles d'apprentissage	Cycles
A	11	32
S	11	847
H	11+17	3757
S+H	11+17	(847 +) 807

Tableau 1 : Nombres moyens de cycles d'apprentissage observés pour les 4 modes d'apprentissage : autonome (A), supervisé (S), hybride (H), supervisé-hybride (S+H). La première colonne du tableau donne le nombre de configurations à apprendre pour chacun des modes.

Nous avons étudié plus précisément 3 transitions : [i a], [u ɔ] et [a y]. Les signaux ont été échantillonnés à 16 kHz. L'extraction des formants a été réalisée à l'aide du CSL-Computerized Speech Lab Model 4300 de la société Kay Elemetric Corp. Après un sous-échantillonnage à 8 kHz, une analyse LPC d'ordre 10 avec une fenêtre d'analyse de 30 ms déplacée de 5 ms, fournit les trajectoires formantiques.

Le *Tableau 2* reprend pour chaque mode d'entraînement et pour chaque transition formantique, le nombre moyen de cycles d'apprentissage nécessaires à l'obtention des trajectoires articulatoires. Le critère d'apprentissage est un critère sur les formants et est donné par les équations 1, 2, 3.

Modes	[i a] (29 pts)	[u ɔ] (46 pts)	[a y] (25 pts)
A	66	58	39
S	49	68	46
H	53	55	40
S+H	45	50	26

Tableau 2 : Nombres moyens de cycles d'apprentissage (optimisation locale) nécessaires à la génération des trajectoires observés pour les 4 modes d'apprentissage: autonome (A), supervisé (S), hybride (H), supervisé-hybride (S+H).

L'extraction des trajectoires articulatoires par le réseau est très rapide quel que soit le mode d'apprentissage réalisé, avec un léger avantage pour le mode supervisé-hybride (*Tableau 2*).

Nous présentons en *Figure 5* et *Figure 6* les configurations articulatoires obtenues respectivement pour deux transitions vocaliques: [i a] et [u ɔ]. Les figures *a, b, c, d* donnent respectivement les configurations obtenues suite à un apprentissage autonome, supervisé, hybride et supervisé-hybride au temps t = 0, T/2 et T ms (où T est la durée de la transition formantique).

Au niveau des configurations articulatoires (*Figure 5* et *Figure 6*), on observe une grande disparité entre les résultats de l'apprentissage autonome et des autres modes. Cette remarque est très bien quantifiée dans l'estimation des volumes moyens où l'apprentissage autonome donne lieu à un volume moyen pratiquement réduit de moitié (*Tableau 3*).

Modes	Volumes moyens (cm^3)
A	50.0
S	92.5
H	90.0
S+H	92.0

Tableau 3 : Volumes moyens (en cm^3) estimés sur l'ensemble du corpus étudié pour les apprentissages autonome (A), supervisé (S), hybride (H) et supervisé-hybride (S+H).

Ceci confirme que, pour l'apprentissage autonome, une contrainte de volume constant doit être intégrée (Soquet, Saerens & Jospa, 1990). Pour les autres modes d'apprentissage, les résultats sont très homogènes. Du point de vue de la forme (*Figure 5* et *Figure 6*) et des volumes moyens (*Tableau 3*) des configurations articulatoires, les modes supervisé, hybride et supervisé-hybride donnent lieu à des résultats forts similaires : les configurations sont acceptables et celles des voyelles de fin de transition sont consistantes avec les résultats publiés (Mrayati 1976; Majid, Boe & Perrier, 1986). Ceci indique que l'utilisation d'un apprentissage supervisé permet d'introduire une connaissance explicite qui permet l'obtention de configurations articulatoires réalistes. La combinaison de cet apprentissage avec un autonome se réalise aisément et fournit des données supplémentaires qui permettent une meilleure gestion par le réseau de données acoustiques relativement éloignées de celles fournies par l'apprentissage supervisé.

CONCLUSION

Par ce travail, nous avons montré qu'un contrôleur neuronal convenablement entraîné, utilisant la méthode variationnelle pour le calcul du Jacobien de la transformation acoustico-articulatoire, permet de générer des transitions articulatoires satisfaisantes.

Nous avons testé plusieurs modes d'apprentissage : autonome, supervisé, hybride et supervisé-hybride. Pour chaque mode, nous avons comparé les temps d'apprentissage et de génération de trajectoires articulatoires. L'apprentissage autonome ne permet généralement pas d'obtenir des configurations articulatoires acceptables sans l'ajout de contraintes explicites (telle une contrainte de volume). Les trois autres modes d'apprentissage font intervenir un apprentissage supervisé, lequel introduit une connaissance articulatoire explicite dans le réseau. Les résultats obtenus de la sorte sont plausibles pour chacun de ces trois derniers modes d'apprentissage. Le mode supervisé-hybride réalise le meilleur compromis. Il combine un temps d'apprentissage acceptable à l'obtention de configurations articulatoires équivalentes à celles obtenues en mode hybride.

REFERENCES

Bailly G., Bach M., Laboissière R. & Olesen M., 1990, "Generation of articulatory trajectories using sequential networks". Proceedings of the First ESCA Workshop on Speech Synthesis, Autrans, pp. 67-70.

BOTHOREL A., SIMON P., WIOLAND F., ZERLING J.-P., 1986, "Cinéradiographie des voyelles et consonnes du français". Institut de Phonétique de Strasbourg.

JOSPA P., 1972, "Forme approchée du conduit vocal déduite des fréquences de résonance. Théorie des perturbations et méthode variationnelle". Actes des 5èmes J.E.P., Lannion 1972. pp 225-261.

JOSPA P., 1994, "Formulation variationnelle du lien acoustico-articulatoire". Actes des 20èmes J.E.P., Trégastel.

JOSPA P., SOQUET A. & SAERENS M., 1992, "Acoustical sensitivity functions and the control of the vocal tract model". Signal Processing VI: Theories and Applications, J. Vandewalle, R. Boite, M. Moonen, A. Oosterlinck (editors), 1992 Elsevier Science Publishers, pp. 171-174.

LABOISSIERE R., SCHWARTZ J.-L. & BAILLY G., 1991, "Control motor for speech skills: A connectionnist approach". Proceedings of the 1990 Connectionnist Models Summer School, Touretzky D., Elman J., Sejnowski T. & Hinton G. (editors), Morgan Kaufmann Publishers, pp. 319-327.

MAJID R., BOE L.J. & PERRIER P., 1986, "Fonctions de sensibilité, modèle articulatoire et voyelles du Français". Actes des 15èmes Journées d'Etude sur la Parole, Aix-en-Provence, pp. 59-63.

MRAYATI M., Contribution aux études sur la production de la parole.; thèse d'Etat ès-Sci. phys. Grenoble (1976).

MRAYATI M., CARRE R. & GUERIN B., 1988, "Distinctive regions and modes: A new theory of speech production". Speech Communication, 7, pp. 257-286.

RAHIM M.G., GOODYEAR C.C., KLEIJN W.B., SCHROETER J. & SONDHI M.M., 1993, "On the use of neural networks in articulatory speech synthesis". J. Acoust. Soc. Am. 93 (2), 1993, pp. 1109-1121.

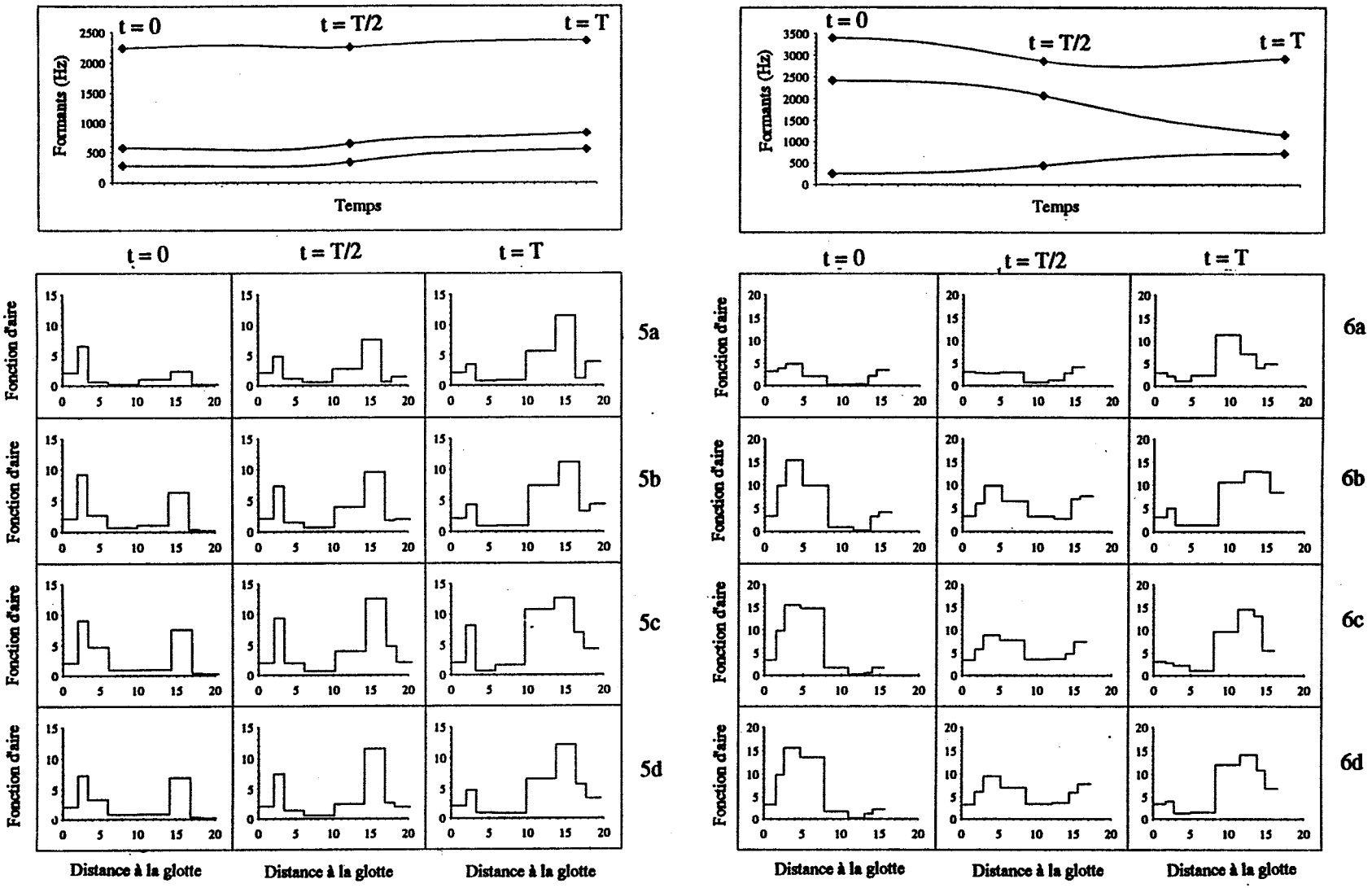
SAERENS M., 1991, "Approche connexioniste de la commande de processus. Application à l'inversion acoustico-articulatoire d'un modèle de conduit vocal, en vue de l'identification du lieu d'articulation des occlusives du Français". Thèse de Doctorat en Sciences Appliquées, Université Libre de Bruxelles, Institut de Phonétique.

SAERENS M., SOQUET A., 1991, "Neural Controller Based on Back-Propagation Algorithm". IEE Proceedings-F, 138(1), pp. 55-62.

SAERENS M., SOQUET A., RENDERS J.-M. & BERSINI H., 1992, "A preliminary comparison between a neural adaptive controller and a model reference adaptive controller". Proceedings of the International Workshop on Neural Networks in Robotics; Kluwer.

SHIRAI K. & KOBAYASHI T., 1991, "Estimation of articulatory motion using neural networks". Journal of Phonetics, 19, pp. 379-385.

SOQUET A., SAERENS M. & JOSPA P., 1990, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model". Proceedings of the ESCA First International Workshop on Speech Synthesis, Autrans, pp. 71-74.



Figures 5-6: Trajectoires formantiques[u ɔ] (Figure 5) et [i a] (Figure 6) mesurées (parole naturelle) et configurations articulatoires correspondantes aux instants t = 0 ms, t = T/2 ms, t= T ms (T: durée de la transition) obtenues par le réseau pour un apprentissage autonome (a), supervisé (b), hybride (c) et supervisé-hybride (d). Les erreurs entre les formants cibles et générés par le réseau ne dépassent pas 30 Hz, 40 Hz et 50 Hz respectivement pour les 1er, 2ème et 3ème formants.

EFFET DE LA CONNAISSANCE DE LA VOYELLE SUR LA PERCEPTION DU BRUIT D'EXPLOSION DES OCCLUSIVES.

LINDA DJEZZAR djezzar@loria.fr
ANNE BONNEAU bonneau@loria.fr
YVES LAPRIE laprie@loria.fr

CRIN-CNRS / INRIA Lorraine
B.P. 239 54506 Vandœuvre-lès-Nancy France

Résumé

Dans cet article, nous présentons une expérience concernant l'effet de la connaissance de la voyelle sur la perception du lieu d'articulation des occlusives sourdes. Les stimuli consistent en des segments de bruit d'explosion extraits de mots monosyllabiques de la base BDSONS.

Cette expérience fait suite à deux expériences déjà publiées [Bonneau 93], que nous résumerons.

L'effet de la voyelle sur les scores de la deuxième expérience étant très significatif, nous nous sommes demandés, si la connaissance de la voyelle pouvait améliorer ces scores. D'où l'élaboration d'une troisième expérience (objet du présent article) dans laquelle les auditeurs sont amenés à identifier l'occlusive tout en connaissant la voyelle qui suit (/i, a, u/).

L'analyse statistique des résultats a révélé une amélioration globale des scores quand la voyelle est connue à l'avance, cette amélioration a concerné les deux occlusives : /k/ (surtout quand la voyelle est /a/) et /p/ (quand la voyelle est /u/).

Mots-clés : perception, lieu d'articulation des occlusives, bruit d'explosion, et effet de la voyelle.

1 INTRODUCTION

Les deux principaux événements acoustiques caractérisant le lieu d'articulation des occlusives sont le bruit d'explosion et les transitions formantiques. Ils ont fait l'objet de nombreux travaux, l'importance relative de chacun d'eux est encore un sujet de discussion. Dans la littérature, il existe peu d'expériences de perception avec des stimuli naturels, portant sur l'étude du pouvoir de discrimination du bruit d'explosion par rapport au lieu d'articulation.

Halle et al. [Halle 57] ont effectué une expérience sur la perception des occlusives /p, t, k/ en position finale, ils ont fait écouter les 20 premières ms du bruit d'explosion. Il ont obtenu, pour les 5 meilleurs auditeurs, les scores suivants : 65, 70, 75, 80 et 96%, leurs résultats accréditent l'hypothèse selon laquelle l'occlusive peut être identifiée indépendamment de la voyelle suivante.

Grâce à plusieurs séries d'expériences, Winitz et al. [Winitz 72] ont conclu que le bruit d'explosion isolé de son contexte vocalique est suffisamment discriminant pour /t/, l'est un peu moins pour /p/ et est insuffisant pour /k/, voir même insignifiant dans le contexte /i/. Les propriétés temporelles et spectrales étant toutes deux présentes, nous ne pouvons pas apprécier la contribution relative de chacune d'elles.

Tekieli et Cullinan [Tekieli 79] ont fait écouter des portions de syllabes /p, t, k, b, d, g/ * /i, I, u, U, æ, ɛ, ɑ,ʌ/ prononcées par un seul locuteur. Pour chaque syllabe, des portions de durée variant de 10 à 150 ms ont été utilisées. Ils conclurent que les indices du lieu d'articulation des 6 occlusives contenus dans les 10 premières ms du bruit d'explosion permettent un score d'identification supérieur à la chance.

L'expérience de Kewley-Port et al. [K.-Port 83] a concerné la perception du bruit d'explosion de /b, d, g/ dans 5 contextes vocaliques /i, e, ɑ, o, u/. Les syllabes sont prononcées par 2 locuteurs. Chacune d'elles a été tronquée en 5 différents points (bruit d'explosion, bruit d'explosion+2 périodes de la voyelle, bruit d'explosion+3 périodes, bruit d'explosion+5 périodes, bruit d'explosion+7 périodes). Les portions de syllabes obtenues varient de 6 à 111ms. Les stimuli issus de la première coupure peuvent inclure la première période de la voyelle suivante (cas de /b/), ou de l'aspiration (cas de /d/ et de /g/) et sont identifiés dans 80% des cas : (b/ et /d/ : 90%, alors que /g/ :70%). Le taux d'identification de /g/ dans le contexte /i/ chute énormément (30%) et il est le plus élevé avec les stimuli (bruit d'explosion+3 périodes) : 60%. Des informations de nature diverse coexistent et notamment de type temporel, donc nous ne pouvons pas, une fois de plus, juger la contribution des propriétés spectrales du bruit d'explosion dans la perception du lieu d'articulation.

De même, Repp et Lin [Repp 89] ont fait écouter des portions de bruit d'explosion de même durée (25,6 ms),

extraits des occlusives /b, d, g/ qui sont chuchotées sans aspiration et sans friction par un seul locuteur. Ils obtiennent un taux d'identification global égal à 51%. Bien que ce score soit supérieur à la chance il est insuffisant ; les stimuli ne sont pas naturels et donc sont dépourvus de certains indices présents en parole naturelle.

Dans l'expérience de Krull [Krull 90], des bruits d'explosion extraits des 4 occlusives du suédois /b, d, d̪t, g/ ont été soumis aux auditeurs. La durée des stimuli est 26 ms. Les scores obtenus sont : /b/ :91%, /d/ :66%, /d̪t/ :69% et /g/ :55%. Les 26 ms peuvent contenir le bruit d'explosion en entier ou le début des transitions formantiques. L'auteur précise que dans la majorité des cas les deux événements sont tous deux présents, que certains stimuli (vélaires) n'incluent que le bruit d'explosion et que d'autres (labiaux) n'incluent que les transitions formantiques.

En résumé, toutes ces expériences montrent que le pouvoir de discrimination du bruit d'explosion (tel que chaque auteur l'a isolé) est important par rapport au lieu d'articulation. Cependant, dans la plupart des expériences, les stimuli incluent toujours, ou bien une part des transitions formantiques (Krull, Kewley-Port), ou bien de l'aspiration (Kewley-Port, Winitz), ou bien les propriétés temporelles (Kewley-Port,Winitz). Il n'y a que l'expérience de Halle (qui concerne seulement le bruit d'explosion en position finale et donc la voyelle n'influence pas beaucoup) et celle de Tekieli et Cullinan (stimuli de 10 ms) où la perception du lieu d'articulation est essentiellement basée sur les propriétés spectrales du bruit d'explosion.

Dans la même optique, nous avons effectué deux séries d'expérience : la première expérience avait pour but de déterminer la durée minimale d'un segment de syllabe CV nécessaire à l'identification du lieu d'articulation de /p, t, k/ [Bonneau 93]. Autrement dit, est-ce que le bruit d'explosion des occlusives sourdes du français en position initiale, isolé de son contexte, est identifiable?

Dans la seconde expérience, nous avons testé le rôle des propriétés spectrales du bruit d'explosion dans la perception du lieu d'articulation dans différents contextes vocaliques sans faire intervenir les propriétés temporelles [Bonneau 93].

Nous proposons maintenant une troisième expérience, afin d'étudier l'effet de la connaissance du contexte sur la perception du bruit d'explosion.

L'originalité de cette contribution, réside dans le fait que les stimuli sont de même durée et sont constitués de telle sorte que seuls la barre d'explosion et un segment de bruit de friction soient présents. Et donc, seules les propriétés spectrales du bruit d'explosion sont testées, précisons néanmoins que l'amplitude n'est pas normalisée.

Ces expériences de perception font partie d'une étude acoustico-perceptive du bruit d'explosion que nous effectuons actuellement. Les connaissances acquises nous serviront à implanter des détecteurs d'indices du bruit d'explosion pour identifier les occlusives.

2 EXPERIENCES PRECEDENTES

Dans ce paragraghe, nous résumons les deux premières expériences qui ont déjà été publiées dans [Bonneau 93], par contre, nous développons l'analyse statistique des résultats de la deuxième expérience, qui à l'époque n'était pas achevée.

2.1 Conditions expérimentales

Les tests se sont déroulés dans une pièce calme. Les sons sont écoutés à travers un casque hi-fi Sennheiser HD 520 qui est connecté à un ordinateur HP modèle 715/ 33. Le chargement et la restitution des sons ont été faits à l'aide d'un éditeur de signal. Ce matériel d'expérimentation est le même pour toutes nos expériences.

2.2 Expérience I : identification de syllabes tronquées

63 syllabes (/p, t, k/ * /i, e, a, y, o, u, ɔ/ * 3 locuteurs), extraites de mots monosyllabiques de BDSONS ont été tronquées en 4 points différents :

i) le premier point de troncation est situé à 2/3 de la durée totale de la voyelle.

ii) le second point est situé à 1/3 de la durée totale de la voyelle.

iii) le troisième point est situé à 20 ms partir du début de la voyelle.

iv) le quatrième point a été effectué au dernier passage par zéro précédant la première période vocalique.

Les stimuli ont été répartis en quatre fichiers selon le point de troncation. Nous avons effectué donc 4 tests, chaque test correspond à l'audition d'un fichier. Les résultats ont montré que la présence d'un segment vocalique, si bref soit-il, permet une identification presque parfaite de l'occlusive. Ils ont confirmé aussi que le bruit d'explosion isolé de son contexte vocalique peut être correctement perçu par les auditeurs.

Les stimuli que nous avons utilisés dans cette expérience ont deux types de propriétés : temporelles et spectrales qui contribuent ensemble dans la perception du lieu d'articulation de l'occlusive. Dans l'expérience suivante, nous avons testé seulement la contribution des propriétés spectrales.

2.3 Expérience II : Identification de bruits d'explosion de même durée

Nous avons construit deux corpus : un pour l'apprentissage et un second pour le test d'identification. Les stimuli qui constituent le corpus d'apprentissage sont différents de ceux du corpus de test.

- **Corpus d'apprentissage**

Il est constitué de 54 stimuli : /p, t, k/ * /i, a, u/ * 3 locuteurs * 2 répétitions.

- **Corpus de test**

Il est constitué de 273 stimuli : (/p, t, k/ * /i, e, ɛ, a, y,

o, œ, Ø, u, ɔ, ɔ̃/ * 5 locuteurs * 2 répétitions) - quelques stimuli manquants.

- **Construction des stimuli**

Les stimuli sont des segments de bruit d'explosion de même durée (entre 20 et 25 ms), ils ne contiennent pas du tout de voyelle ni de transitions formantiques. Afin d'obtenir de tels segments, nous en avons tronqué certains et allongé d'autres:

- Les stimuli /t/ sont en général de durée comprise dans l'intervalle requis, sauf quelques uns qu'on a dû tronquer.
- La plupart des stimuli /k/ ont été tronqués et amortis à droite afin d'atténuer la coupure.
- La plupart des stimuli /p/ sont allongés ; nous avons dupliqué la fin du bruit de friction autant de fois que c'est nécessaire pour atteindre la durée 25ms. Nous avons veillé à que tous les stimuli allongés soient perceptiblement corrects.

2.3.1 Analyse des résultats.

Les pourcentages des bonnes réponses et des confusions sont présentés dans la table I.

Il est clair que les scores sont très élevés quelle que soit l'occlusive et dans n'importe quel contexte vocalique; ils sont nettement supérieurs à la chance. Nous avons d'abord procédé à une analyse de variance (ANOVA) à trois paramètres (Audition*Lieu d'articulation*Voyelle) ; le paramètre Audition prend 2 valeurs puisque chaque stimulus est écouté 2 fois par chaque auditeur. Et ensuite, nous avons utilisé le test de Scheffé comme test de comparaison multiple. Comme le nombre de stimuli concernant les voyelles /œ, e, ɔ̃/ est insuffisant, nous les avons écartées de l'analyse de variance.

Les résultats de l'analyse ANOVA montrent que le paramètre Audition a un effet significatif sur les scores [$F(1,446)=5.75$, p.c.< 0.05]. En effet, une amélioration a été observée au cours du dernier jour, due probablement à la familiarisation des auditeurs avec les sons et avec les conditions de l'expérience.

En ce qui concerne le lieu d'articulation, les différences de performance se sont révélées significatives [$F(2,446)=4.66$, p.c<0.01]. Les résultats du test de Scheffé nous affirment que le bruit d'explosion de /p/ est mieux perçu que celui des deux autres occlusives. Pourquoi /p/ est-t-il le mieux reconnu? Trois réponses non exclusives sont possibles :

. Il peut être mieux reconnu grâce au «traitement temporel» qu'il a subi ; nous rappelons que seuls les bruits d'explosion de /p/ ont été allongés.

. ou à son amplitude qui est, en général, relativement faible par rapport à celle des deux autres occlusives (sauf peut-être quand il est suivi de voyelles fermées).

. ou plutôt à sa forme spectrale.

Afin de pouvoir répondre d'une manière beaucoup plus précise, nous devons examiner tous les stimuli /p/, séparer ceux-ci en deux classes : la classe des /p/ qui sont bien identifiés et la classe des /p/ qui sont mal identifiés. Puis, dans chacune des classes, voir s'ils ont subi le traitement temporel ou non, calculer leur amplitude et déterminer les caractéristiques de leur forme spectrale. Ce travail est en cours, nous n'avons pas encore obtenu de résultats.

Table I : Pourcentages des réponses et des confusions dans l'expérience II.

	Réponses		
Stimuli	P	T	K
/pi/	**89**	6	5
/ti/	0	**90**	10
/ki/	0	28	**72**
/pe/	**89**	9	2
/te/	2	**94**	4
/pɛ/	**87**	11	2
/tɛ/	1	**85**	14
/kɛ/	1	28	**71**
/pa/	**87**	10	3
/ta/	15	**79**	6
/ka/	0	28	**72**
/py/	**83**	7	10
/ty/	0	**72**	28
/ky/	3	13	**84**
/pØ/	**97**	2	1
/tØ/	0	**83**	17
/kØ/	2	6	**92**
/tœ/	0	**98**	2
/kœ/	1	12	**87**
/pu/	**85**	6	9
/tu/	0	**97**	3
/ku/	0	02	**98**
/po/	**91**	8	1
/to/	1	**86**	13
/ko/	0	1	**99**
/pɔ/	**93**	4	3
/tɔ/	8	**83**	9
/kɔ/	1	1	**98**
/pɔ̃/	**92**	6	2
/tɔ̃/	8	**77**	15
/kɔ̃/	2	0	**98**

Un troisième effet qui s'est révélé très significatif, est celui de la voyelle [$F(7,446) = 10.47$, p.c. < 0.001]. Nous rappelons que seules les voyelles {u, o, ɔ, Ø, i, ɛ, y, a} ont été prises en compte dans l'analyse. Le test de Scheffé fait ressortir deux groupes de voyelles :

- les voyelles devant lesquelles les stimuli sont les mieux identifiés : {u, o}.
- les voyelles devant lesquelles les stimuli sont les moins bien identifiés : {a, y}.

Nous avons examiné toutes les interactions. Il n'y a que l'interaction (Lieu d'articulation * voyelle) qui s'est révélée significative [$F(14,446)=6.86$, p.c.<0.01]. Le test de Scheffé fait ressortir deux classes de syllabes :

- les syllabes {/ki/, /kɛ/, /ka/, /pu/} où les bruits d'explo-

sion sont les moins bien perçus (par rapport au lieu d'articulation et à la voyelle),
- les syllabes {/ko/, /kɔ/, /ku/, /ti/} où les bruits d'explosion sont les mieux perçus (par rapport au lieu d'articulation et à la voyelle).

Certaines combinaisons (occlusive*voyelle) semblent favoriser la perception du lieu d'articulation de l'occlusive alors que d'autres, au contraire, la dégradent :
- /k/ suivi des voyelles /u, o, ɔ/ est le mieux perçu, ceci peut être dû à certaines affinités que possède /k/ avec ces voyelles ; ils ont le même lieu d'articulation (vélaire).
- Pareillement, le fait que la voyelle /i/ est palatale et fermée et /t/ est apico-dental semble favoriser la perception de dernier.
- Par contre, /k/ suivi des voyelles /i, a, ɛ/ est confondu avec /t/. La confusion dans le contexte /i/ est pratiquement rencontrée dans tous les travaux cités dans l'introduction. Blumstein et Stevens [Blumstein 79] l'expliquent par le fait que les hautes fréquences dominent le spectre du bruit d'explosion de /k/. Nous avons constaté le même phénomène pour les voyelles /a, ɛ/ ; les trois voyelles sont antérieures et non arrondies dans ce contexe.
- /p/ suivi de / u/ est confondu avec /k/; les basses fréquences des formants F2 et F3 renforcent l'énergie dans les régions basses et provoquent ainsi une réponse /k/.

2.4 Discussion

Les scores d'identification qui se sont révélés très élevés quelle que soit la consonne et quelle que soit la voyelle qui suit, confirment que la forme générale du spectre du bruit d'explosion (barre d'explosion+segment de bruit de friction) a un rôle important dans la perception du lieu d'articulation des occlusives sourdes. L'analyse statistique révèle que les scores varient en fonction de la voyelle. Blumstein et Stevens [Blumstein 79] ont noté aussi un effet significatif de la voyelle et ils l'ont expliqué par le phénomène de coarticulation dû à l'anticipation de la masse de la langue vers le lieu d'articulation de la voyelle. Les expériences de Winitz et al. [Winitz 72], et de Cullinan et Tekieli [Tekieli 79] ont montré que la voyelle peut-être identifiée, avec des scores variables mais supérieurs à la chance, à partir du bruit d'explosion seulement (avec aspiration). Suomi [Suomi 85] a conclu que pour un lieu d'articulation donné la forme spectrale du bruit d'explosion est dépendante de la voyelle qui suit. Donc, nos résultats concordent bien avec ceux des auteurs cités, et nous confirment que la perception du lieu d'articulation de l'occlusive est influencée par la voyelle qui suit.

Cette influence étant très significative, nous nous sommes demandés si la connaissance de la voyelle pouvait avoir un effet sur la perception du lieu d'articulation de la consonne. D'où l'élaboration de l'expérience III.

3 EXPERIENCE III : IDENTIFICATION DU BRUIT D'EXPLOSION DANS UN CONTEXTE CONNU A L'AVANCE.

Dans cette expérience, nous élaborons un test par voyelle. Nous nous sommes restreint aux trois voyelles /a, i, u/.

3.1 Corpus

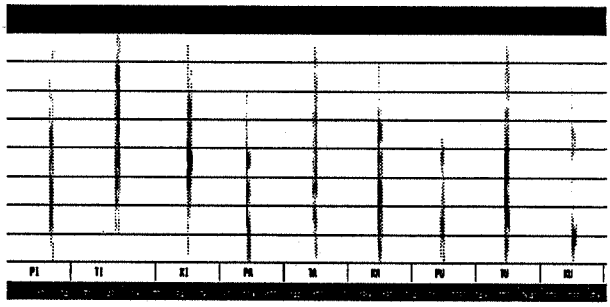
- **Corpus d'apprentissage**

Nous avons repris le même corpus que celui utilisé pour l'apprentissage dans l'expérience II. Mais nous l'avons organisé de manière à obtenir un fichier par voyelle /i, a, u/. Dans chacun des trois fichiers, les stimuli sont regroupés par occlusive.

- **Corpus de test**

Il est extrait du corpus de test de l'expérience II et est donc composé de 9 syllabes (/p, t, k/ * /i, a, u/) prononcées 2 fois par 5 locuteurs, il y a donc 10 items différents par syllabe. Les 90 items obtenus sont répartis en 3 fichiers, un fichier par voyelle. Chacun d'eux est organisé sous forme de 3 blocs de 10 items. Les items sont séparés par un silence de 3 secondes. Nous présentons ci-dessous le spectrogramme de quelques stimuli.

Figure I : Spectrogramme de quelques stimuli de même durée.



3.2 Sujets

8 auditeurs ont participé à cette expérience, les mêmes que ceux de l'expérience II qui ont reçu un apprentissage complet.

3.3 Procédure

L'expérience est organisée en trois séances, à des jours différents mais successifs :

3.3.1 Séance 1

C'est une séance d'apprentissage intense où l'auditeur fait les quatre tests de l'expérience I puis l'apprentissage de l'expérience II.

3.3.2 Séance 2

Elle est composée de trois étapes, une étape par voyelle X, chaque étape se déroule en trois phases : apprentissage, test puis pause. A chaque étape, l'identité de la voyelle X est donnée à l'auditeur. Afin que l'identification de l'occlusive ne soit pas biaisée par l'ordre des étapes (l'ordre des voyelles), nous avons utilisé les 6 ordres possibles, puisque nous avons 3 voyelles.

- **Apprentissage**

1. L'auditeur écoute 3 fois le fichier d'apprentissage

concernant la voyelle X avec les réponses fournies sur une feuille.

2. Il fait le test sur le même corpus mais cette fois-ci trié aléatoirement.

3. Nous corrigeons ensemble les erreurs, puis il reécoute le corpus 3 fois en se concentrant sur ses erreurs.

• **Test d'identification**

Le corpus de test de la voyelle X composé de 30 items est présenté à l'auditeur, il doit reconnaître chaque item parmi {/p/, /t/, /k/ ou «aucune réponse»} sachant qu'il connaît la voyelle. La possibilité «aucune réponse» n'a pas été utilisée.

Après le test, l'auditeur se repose pendant un quart d'heure. Avant de commencer l'étape suivante avec une autre voyelle.

3.3.3 Séance 3

La troisième séance se déroule le jour suivant, exactement de la même manière que la seconde sauf que tous les stimuli, que ce soient ceux de l'apprentissage ou du test sont présentés dans un ordre aléatoire différent. De cette manière nous obtenons deux auditions pour chaque stimulus.

3.4 Résultats et discussion

Les résultats sont présentés dans la table II. Ils restent très élevés et sont beaucoup plus homogènes.

Table II : Pourcentages des réponses et des confusions dans l'expérience III.

	Réponses		
Stimuli	P	T	K
/pi/	**92**	1	7
/ti/	0	**86**	14
/ki/	0	22	**78**
/pa/	**87**	8	5
/ta/	8	**79**	13
/ka/	1	9	**90**
/pu/	**93**	3	4
/tu/	0	**99**	1
/ku/	1	1	**98**

Les tests de l'analyse ANOVA (Audition*Lieu d'articulation*Voyelle) indiquent que seul le paramètre Voyelle a un effet sur les scores [F(2,166)=13.28, p.c.<0.001], le test de Scheffé confirme que les bruits d'explosion suivis de /u/ sont les mieux identifiés. La seule interaction qui s'est révélée significative est l'interaction (Lieu d'articulation*Voyelle) [F(4,166) = 7.67, p.c.< 0.001]. Le test de comparaison multiple fait ressortir une seule syllabe /ki/ dont le bruit d'explosion est le moins bien identifié (par rapport au lieu d'articulation et à la voyelle).

4 COMPARAISON ENTRE EXPERIENCE II ET EXPERIENCE III

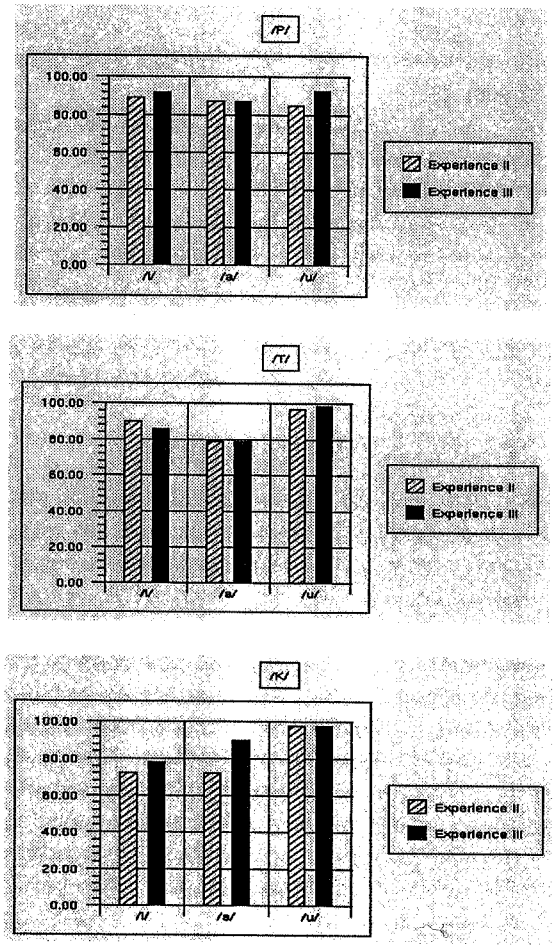
La figure II illustre les variations des scores d'identification pour chaque occlusive en fonction de la voyelle et de l'expérience. A première vue, nous constatons une certaine amélioration des scores des occlusives /k/ et /p/ dans l'expérience III (voyelle connue à l'avance). Par contre, les scores de l'occlusive /t/ restent constants avec une légère diminution dans le contexte /i/.

Afin de vérifier si ces écarts sont significatifs ou non, nous avons procédé à une analyse ANOVA à 4 paramètres (Expérience*Audition*Lieu d'articulation*Voyelle). Le paramètre Expérience prend 2 valeurs : Expérience II et Expérience III.

L'analyse révèle que les scores de l'expérience III sont globalement meilleurs que ceux de l'expérience II [F(1,336)=7.05, p.c.<0.01) ; les performances de l'expérience III ont augmenté de 3.7% par rapport à celles de l'expérience II.

L'effet de la voyelle sur les scores est très significatif [F(2,336)= 28.83, p.c.<0.001], le test de Scheffé indique que les scores concernant la voyelle /a/ sont les moins bons, les deux expériences confondues. Par contre ceux concernant la voyelle /u/ sont les meilleurs.

Figure II : Variation des pourcentages d'identification en fonction de la voyelle et de l'expérience pour chaque occlusive.



L'effet du lieu d'articulation est aussi significatif [F(2,336)=5.50, p.c.<0.01] ; le test de Scheffé montre

que /k/ est le moins bien perçu, les deux expériences confondues.

L'interaction double (lieu d'articulation * voyelle) reste très significative [F(4,336)=14.40, p.c.<0.001] ; le test de Scheffé indique que le score de /k/ suivi de /u/ est le plus élevé, par contre, le score de /p/ suivi de /u/ et celui de /k/ suivi de /i/ sont, par rapport au lieu d'articulation et à la voyelle, les moins élevés.

Le paramètre Expérience n'a pas d'effet ni sur l'audition ni sur la voyelle, par contre, il a un effet peu significatif sur le lieu d'articulation [F(2,336)=1.90; p.c.< 0.25], le test de Scheffé indique une amélioration du score d'identification de /k/. Un autre effet qui s'est révélé plus significatif est celui de l'interaction triple (Expérience * Lieu d'articulation * Voyelle) [F(4,336)=2.42, p.c.<0.05]. Le test de Scheffé confirme que le score de /ka/ est celui qui a le plus augmenté, que l'augmentation du score de /pu/ est peu significative.

Repp et Lin [Repp 89] ont trouvé que les scores de l'expérience où la voyelle est connue à l'avance sont globalement meilleurs que ceux de l'expérience où la voyelle est inconnue, ce qui concorde bien avec notre premier résultat (effet de l'expérience). Cependant ils affirment que la barre d'explosion devant les voyelles d'arrière est moins bien identifiée quand la voyelle est connue à l'avance notamment pour /b/. Nous n'avons pas trouvé un pareil résultat ; nos scores, concernant la voyelle /u/, sont restés constants pour /k/ et /t/ et ont augmenté pour /p/.

Nous pouvons conclure que le fait de connaître la voyelle qui suit peut améliorer la perception de :

• /k/ surtout quand cette voyelle est /a/ ; nous constatons que le score de /ki/ est le même que celui de /ka/ quand la voyelle est ignorée (72%), cependant il n'a augmenté que de 6% quand la voyelle est connue, ce qui n'est pas significatif d'après les tests de l'analyse ANOVA. En ce qui concerne /ku/, il est initialement très bon et il le reste.

• /p/ quand cette voyelle est /u/.

5 CONCLUSION GENERALE

Les expériences que nous avons effectuées ont montré que le bruit d'explosion isolé de son contexte peut être correctement perçu par l'oreille humaine. En neutralisant la contribution des propriétés temporelles, les scores sont élevés avec une amélioration notable du taux d'identification de /p/. L'effet de la voyelle qui s'est révélé très significatif suggère que la perception du lieu d'articulation des occlusives sourdes, qui est basée essentiellement sur la forme spectrale du bruit d'explosion, est fortement influencée par la voyelle. Devant les voyelles d'arrière /u, o, ɔ/, les occlusives sont mieux perçues, par contre, devant les voyelles /a, y/, elles sont moins bien perçues. La connaissance du contexte vocalique améliore globalement les scores d'identification notamment pour /k/ dans le contexte /a/ et pour /p/ dans le contexte /u/.

Nous avons présenté une étude perceptive portant sur le pouvoir de discrimination du bruit d'explosion par rapport au lieu d'articulation. Actuellement, nous analysons acoustiquement les stimuli utilisés dans ces expériences afin d'étudier la corrélation pouvant exister entre les résultats perceptifs et les résultats acoustiques.

La finalité étant d'extraire suffisamment de connaissances, sur les plans perceptif et acoustique afin d'implanter des détecteurs d'indices robustes.

Références

[Blumstein 79] S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Am., 66(4):1001-1017, 1979.

[Bonneau 93] A. Bonneau, L. Djezzar and Y. Laprie. Perception of French stop bursts, implications for stop identification. EuroSpeech'93, Berlin, vol. I: 693-696, 1993.

[Halle 57] M. Halle, G. W. Hughes and J.-P. Radley. Acoustic properties of stop consonants. J. Acoust. Soc. Am., 29(1):107-116, 1957.

[K.-Port 83] D. Kewley-Port, D. B. Pisoni and M. Studdert-Kennedy, Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. J. Acoust. Soc. Am., 73(5):1779-1793, 1983.

[Krull 90] D. Krull. Relating acoustic properties to perceptual responses: A study of Swedish voiced stops. J. Acoust. Soc. Am., 88(6):2557-2570, 1990.

[Repp 89] B. H. Repp and H.-B. Lin. Acoustic properties and perception of stop consonant release transients. J. Acoust. Soc. Am., 85(1):379-396, 1989.

[Suomi 85] K. Suomi. The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables. J. Phon. 13, 267-285, 1985.

[Tekieli 79] M. E. Tekieli and W. L. Cullinan. The perception of temporally segmented vowels and consonant-vowel syllables. J. Speech Hear. Res. 22, 103-121, 1979.

[Winitz 72] H. Winitz, M.E. Scheib and J. A. Reeds. Identification of Stops and Vowels for the Burst Portion of /p,t,k/ Isolated from Conversational Speech. J. Acoust. Soc. Am., 51(4):1309-1317, 1972.

MODÉLISATION DES MOUVEMENTS ARTICULATOIRES PAR LA MÉTHODE DE LA MLPC ET SES EFFETS SUR LES CARACTÉRISTIQUES ACOUSTIQUES

Soumya BOUABANA & Shinji MAEDA

Télécom Paris - Département Signal & CNRS-URA 820 46, Rue Barrault 75634 Paris Cedex 13, FRANCE
e-mail : soumya@sig.enst.fr

Résumé

Dans un système de production de la parole, le signal de la commande est-il organisé en fonction de critères phonémiques, syllabiques ou dépend-il d'une entité phonétique à définir? Dans le but de répondre à cette question, nous avons effectué une analyse-synthèse des mouvements articulatoires du conduit vocal. La méthode utilisée est basée sur la technique standard de la LPC multi-impulsionnelle [1]. Les résultats obtenus à l'aide de cette méthode vont à l'encontre de la traditionnelle théorie qui suppose que l'entrée du système est organisée en fonction de critères phonémiques et soutiennent l'hypothèse que le geste élémentaire dans la production de la parole correspond à un geste syllabique. D'autre part, la projection des résultats de la synthèse dans le domaine acoustique s'avère être un bon outil pour vérifier les caractéristiques formantiques.

Mots-clés: Communication parlée, mouvement articulatoire, modélisation, commandes articulatoires, profil sagittal, fonction de transfert.

1. Introduction

La parole est communément vue comme le résultat de la concaténation de segments phonétiques. Cependant, la question sur la nature de ces segments reste toujours posée. Les divers travaux menés sur la synthèse segmentale se focalisent principalement sur quatre axes : la conversion phonèmes/parole à partir de règles [2], la concaténation de diphones [3], la concaténation de di-syllabes [4] et la synthèse par polysons [5].
L'objectif premier de ce travail est de contribuer à la discussion sur l'unité phonétique en partant de la modélisation des mouvements articulatoires du conduit vocal par la méthode de la LPC multi-impulsionnelle (MLPC). On suppose, en effet, qu'on peut décrire la variation temporelle de chacun des paramètres articulatoires par un modèle source-filtre. Pour chaque modèle, la source est décrite par un train d'impulsions et le filtre est décomposé en deux sous-systèmes g et h invariants dans le temps et placés en cascade comme le montre la figure 1.

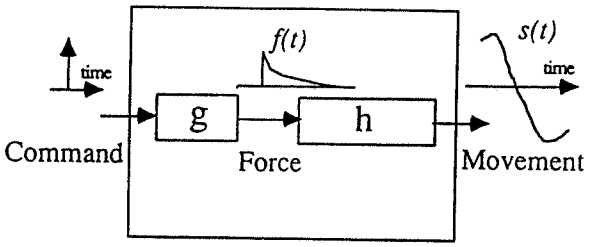


Figure 1. *Modélisation de l'évolution temporelle des paramètres articulatoires de la langue.*

L'avancement de notre étude nous a conduit à des résultats concernant d'une part, l'effet distinctif des paramètres articulatoires de la langue au niveau acoustique et d'autre part, la relation entre les phénomènes articulatoires et les caractéristiques acoustiques.

2. Données

Notre étude porte sur l'analyse du mouvement de la langue durant la production de la parole. Cet articulateur, qui constitue la partie la plus mobile du système supraglottique de l'appareil vocal, participe très activement à la production. Maeda [6] a décrit les variations temporelles du profil de la langue à l'aide de 4 composantes linéaires extraites de l'analyse factorielle appliquée à des données radiofilms [7] : une composante extrinsèque, la position de la mâchoire jw (fermé/ouvert); trois composantes intrinsèques, la position du corps de la langue tp (antérieur/postérieur), la forme du corps de la

langue ts (arquée/plate) et la position de l'apex de la langue tt (haute/basse). Pour la partie labiale et le larynx, trois paramètres articulatoires supplémentaires ont été utilisés. Deux paramètres intrinsèques aux lèvres qui sont la protrusion du pavillon labial lp et la hauteur frontale de l'aperture des lèvres lh. Le troisième paramètre lx décrit le mouvement du larynx. Ces composantes reconstituent la forme du conduit vocal avec une précision suffisante pour calculer les caractéristiques acoustiques.
Nous disposons au total de 10 phrases françaises prononcées par deux locuteurs féminins (PB et DF) et enregistrées à une fréquence de 50 images/sec. A partir de données radiofilms numériques, les valeurs des paramètres sont calculées trame par trame pour fournir la variation temporelle des paramètres (mouvement). Seules les images correspondant à la parole ont été numérisées. Ceci se traduit par la présence de discontinuités au début et à la fin des mouvements. Afin d'atténuer l'effet de bord, les données lissées sont filtrées par une fenêtre de Hamming. La courbe en trait plein de la figure 4 illustre un exemple de l'évolution temporelle du paramètre tp pour la phrase " Une réponse ambiguë".

3. Analyse et synthèse de mouvements par MLPC

3.1. Identification du filtre

Le sous-système g (défini dans la figure 1) du modèle MLPC est un filtre de dé-accentuation. Il est obtenu par un filtre intégrateur du premier ordre qui dépend d'un coefficient α. En synthèse, ce filtre permet la mise en forme du signal d'excitation. En phase d'analyse, le filtre inverse $\mathbf{g}^{-1}$ est appelé filtre de pré-accentuation. Son rôle est d'applatir la pente spectrale du signal d'observation afin de faciliter la détection de pôles. Le coefficient α a été déterminé empiriquement de façon que l'énergie de l'erreur entre le signal synthétisé et le signal observé soit faible. Cette erreur est obtenue pour $\alpha = 0.5$.

Pour l'identification du filtre h, dont l'ordre p est inconnu, on est d'abord confronté à la détermination de la valeur p. Dans le cas d'un modèle AR, Akaïke a proposé un critère pour déterminer l'ordre du filtre : Akaïke Information Criterion [8]. Le critère $AIC(p)$ est défini

$$AIC(p) \triangleq \log(\nabla_p^*/R_0) + 2p/kN,$$

où N, ∇_p^*, R_0 et k sont respectivement le nombre d' échantillons du signal, l'énergie de l'erreur de prédiction pour p pôles, la variance du signal et le coefficient de compensation par fenêtrage ($k = 0.375$ dans le cas de Hamming). La valeur optimum p est donnée par le minimum de $AIC(p)$. Le tracé de la fonction $AIC(p)$, pour les quatre paramètres sur l'ensemble des dix phrases, nous a conduit à choisir $p = 2$. Cette valeur est obtenue dans 90% des cas. Une fois l'ordre du filtre h déterminé, l'identification des coefficients $a_i, i = 1, .., p$ est effectuée par la technique de la LPC.
Pour chaque paramètre articulatoire, dix paires de pôles conjugués ont été ainsi calculées pour les dix phrases et ont été ensuite reportées sur le cercle unité. Pour chaque paramètre, deux nuages de points compacts apparaissent de part et d'autre de l'axe réel positif. Ce résultat est présenté par la figure 2 dans le cas du paramètre tp. Pour avoir les systèmes les plus représentatifs de l'invariance temporelle, les quatre filtres de synthèse correspondant aux quatre paramètres articulatoires sont construits à partir des centres de gravité des nuages de points respectifs.

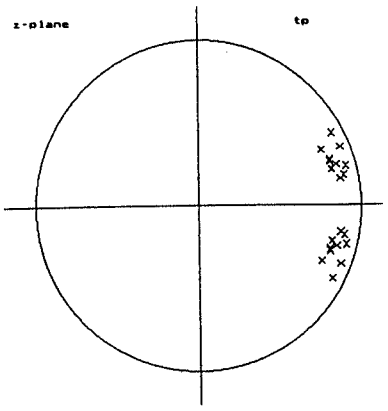


Figure 2. *Position des paires de pôles conjugués pour les dix phrases.*

Comme le montre la figure 3, les réponses impulsionnelles de ces quatre filtres AR présentent une faible oscillation amortie. Le signal de commande étant une séquence

d'impulsions, ces réponses impulsionnelles sont considérées comme le geste élémentaire qui, par superposition, donne naissance aux mouvements articulatoires. Il est important de remarquer que la durée effective de chacune des réponses impulsionnelles est comprise entre 140 et 200ms. Cette durée est pratiquement égale à 180ms, la durée moyenne de la syllabe française .

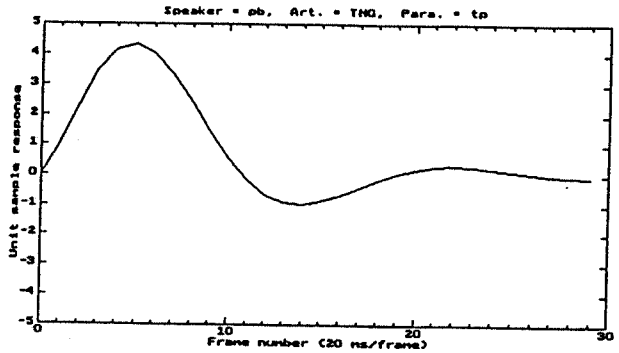


Figure 3. *Réponse impulsionnelle.*

3.2. Interprétation physique du filtre

Dans le paragraphe précédant nous avons décrit le modèle de synthèse à l'aide de deux filtres. Dans la chaîne de production de la parole chez l'homme, on distingue deux systèmes : un système physiologique et un système dynamique (mécanique). Le problème est d'établir une analogie entre notre modèle et la chaîne de production.
Nos premiers travaux nous ont conduit à choisir le filtre AR **h** d'ordre 2 comme système dynamique (mécanique). En effet, le mouvement articulatoire est souvent considéré comme un mouvement dynamique caractérisé par une masse M, une résistance mécanique R et un coefficient de raideur K. Il est décrit par l'équation de Newton :

$$M\ddot{s}(t) + R\dot{s}(t) + Ks(t) = f(t),$$

où $f(t)$ est la force d'excitation [9, 10]. Le filtre de dé-accentuation **g**, fortement amorti, correspond alors au système physiologique. Il consiste à transformer la séquence d'impulsions en une forme d'onde exponentielle qui représente la force d'excitation du système dynamique.
Toutefois, pour modéliser le mouvement de la main durant l'écriture, Yasuhara a identifié deux systèmes $\mathbf{g_m}$ et $\mathbf{h_m}$ [11]. Il a associé le système oscillant $\mathbf{g_m}$ (composé par la paire de pôles complexes conjugués) au système physiologique, et le système fortement amorti $\mathbf{h_m}$ au mouvement dynamique de la main. Son choix est basé sur le calcul de la valeur du coefficient de qualité $Q = f/B$, où f et B sont respectivement la fréquence de résonance et la largeur de bande. Le coefficient Q qui correspond au pôle réel est nul. Il engendre donc le système le plus amorti, c'est-à-dire, le système mécanique.
Pour comparer nos résultats à ceux obtenus par Yasuhara, nous avons calculé le coefficient de qualité pour le système $\mathbf{g_m}$ et pour le système **h** correspondant aux paramètres **jw**, **tp**, **ts** et **tt**. Ainsi que le montre le tableau 1, les valeurs obtenues dans les deux cas sont comparables. Cependant, il est important de noter que les résultats du tableau 1 correspondent au système physiologique dans le cas du mouvement de la main. Si on adapte l'interprétation de Yasuhara à nos deux sous-systèmes, notre filtre de pré-accentuation **g** est alors le système mécanique (et donc visco-élastique) et **h** est le système physiologique représenté par le filtre d'ordre $p = 2$.

articulateur	$f_{(Hz)}$	$B_{(Hz)}$	Q
main	3.42	1.20	2.78
jw	2.47	3.71	0.80
tp	2.46	2.02	1.25
ts	2.60	2.16	1.25
tt	3.16	4.05	0.80

Tableau 1. *Comparaison des fréquences naturelles, des bandes passantes et des coefficients de qualité obtenus pour le mouvement de la main durant l'écriture et celui de la langue durant la parole.*

3.3. Détermination du modèle d'excitation

Supposons que $\hat{s}_m(n)$ est la sortie du synthétiseur AR excité par une séquence $v_m(n)$ qui n'est pas nulle pour m instants fini. L'estimation de $v_m(n)$ est choisie de manière à minimiser

$$\sum_{n=1}^{N}(|s(n) - \hat{s}_m(n)|^2).$$

Une solution sub-optimale, mais efficace, est itérative sur m. Elle utilise

$$e_{m-1}(n) \triangleq s(n) - \hat{s}_{m-1}(n)$$

pour déterminer la $m^{ème}$ position. L'amplitude des impulsions déjà déterminées est recalculée à chaque nouvelle itération [1].

Le mouvement synthétisé en fonction du nombre d'impulsions de la commande d'entrée est présenté dans la figure 4.

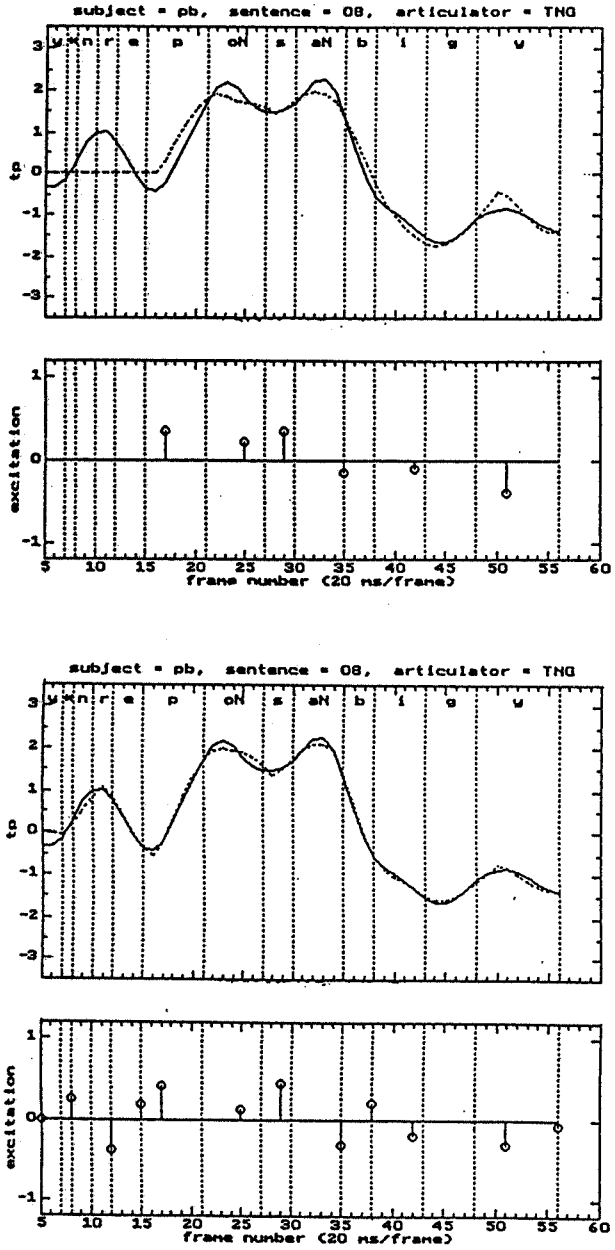


Figure 4. *Évolution temporelle du paramètre* tp *observé (trait plein) et synthétisé (pointillé). Nombre d'impulsions : $m = 6$ (nombre de syllabes) et $m = 12$ (nombre de phonèmes).*

Cette figure montre que la différence entre les mouvements observé et reconstitué diminue lorsque le nombre d'impulsions augmente. Afin d'élucider ce résultat, nous avons calculé l'énergie de l'erreur entre les deux signaux en fonction du nombre d'impulsions m :

$$E(m) \triangleq \left[\sum_{n=1}^{N} |e_m(n)|^2\right] / \left[\sum_{n=1}^{N} y(n)^2\right].$$

Ainsi que le montre l'exemple de la figure 5, nous avons observé que l'erreur diminue rapidement jusqu'à un nombre d'impulsions égal au nombre de syllabes dans la phrase, en l'occurence $m = 6$. Si bien que la pente $P1$ de cette portion de la courbe coupe l'axe des abscisses au point m égal à 6. Au delà de cette valeur et jusqu'à m égal au nombre de phonèmes ($m = 12$), la pente $P2$ de l'erreur devient relativement faible. Dans cette partie du graphe l'énergie de l'erreur est comprise entre 10% et 30% de l'énergie du signal observé. Une saturation est observée lorsque m est supérieur au nombre de phonèmes. Remarquons que ces résultats sont obtenus dans 90% des cas. Pour les phrases où la synthèse d'un paramètre ne vérifie pas les résultats cités ci-dessus, on a observé qu'un des autres paramètres présente une courbe d'erreur dont les pentes $P1$ et $P2$ sont relativement plus rapides, donc nécessite un nombres d'impulsions plus faible. Ainsi, nous sommes tentés de supposer que le nombre globale de commandes pour une phrase est prédéfini.

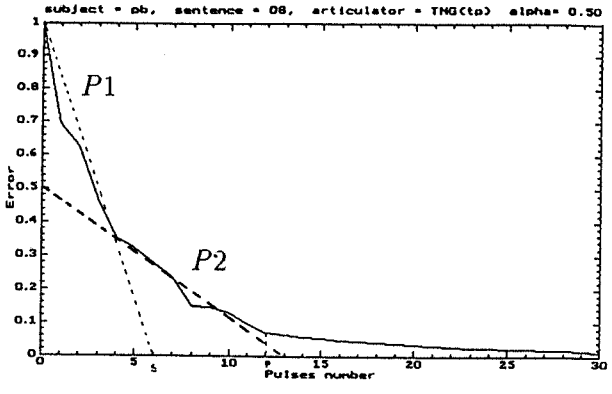


Figure 5. *Énergie de l'erreur de synthèse.*

4. Calcul des fréquences des formants

Une fois les paramètres articulatoires de la langue synthétisés, la deuxième étape de notre travail a été d'examiner les conséquences acoustiques de cette synthèse. Pour ce faire, nous avons d'abord calculé la coupe sagittale du conduit vocal. La forme vectorielle z de la langue

à chaque trame est reconstituée en appliquant une somme pondérée linéaire sur les quatre composantes articulatoires (**jw**, **tp**, **ts**, et **tt**) du mouvement de la langue. Pour aboutir à la forme sagittale correspondante, le vecteur **z** est ensuite projeté sur un système de coordonnées semi-polaires dont la référence est fixée par rapport au maxillaire. Les sections labiale et du larynx sont modélisées par deux tuyaux elliptiques uniformes. Le premier, vers les lèvres, est calculé en fonction des composantes labiales **lp**, **lh** et **jw**. Le second, vers le larynx, est spécifié par sa position **lx** et **jw**. Ces deux tuyaux sont connectés aux deux extrémités du conduit vocal principal.

Puis, nous avons calculé la fonction d'aire. Le passage "coupe sagittale-fonction d'aire", pour une configuration donnée du conduit vocal, est basée sur la formule $A = \alpha x^{\beta}$, où A et x sont respectivement l'aire de la section et la dimension sagittale. Les coefficients α et β dépendent de la position de la grille considérée dans le système semi-polaire.

Une fois la fonction d'aire déterminée, le calcul des fréquences des formants est effectué par un modèle décrivant une ligne de transmission [12]. Ce modèle comporte une source de pression d'air constante, une section étroite variable dans le temps représentant la glotte et un tube correspondant au conduit vocal couplé à la cavité nasale. Le conduit vocal est discrétisé dans l'espace. Il comporte M tubes de longueur variable, de la glotte aux lèvres. Chaque tube est modélisé par un circuit électrique qui exprime les lois acoustiques dans un tuyau non rigide prenant en compte l'effet de la charge acoustique dûe au rayonnement aux lèvres.

Dans le présent article, nous nous intéressons à l'évolution des fréquences des formants des voyelles orales et nasales dans les phrases étudiées en fonction du nombre d'impulsions (sans tenir compte de l'effet du conduit nasal).

5. Rapport entre articulatoire et acoustique

En utilisant la méthode de calcul des formants décrite ci-dessus, nous avons constitué une base de données des valeurs des quatre premiers formants des voyelles des dix phrases françaises à partir des sept paramètres articulatoires observés. Nous avons ensuite effectué le même calcul pour chacune des fonctions d'aire obtenue à partir d'un seul paramètre articulatoire synthétisé de la langue. Les six autres paramètres nécessaires pour la reconstitution de la forme du conduit vocal sont les paramètres observés.

La figure 6 illustre un exemple des formants observés (trait plein) et des formants calculés (pointillé) en fonction du nombre d'impulsions du signal d'excitation pour le paramètre **tp** .

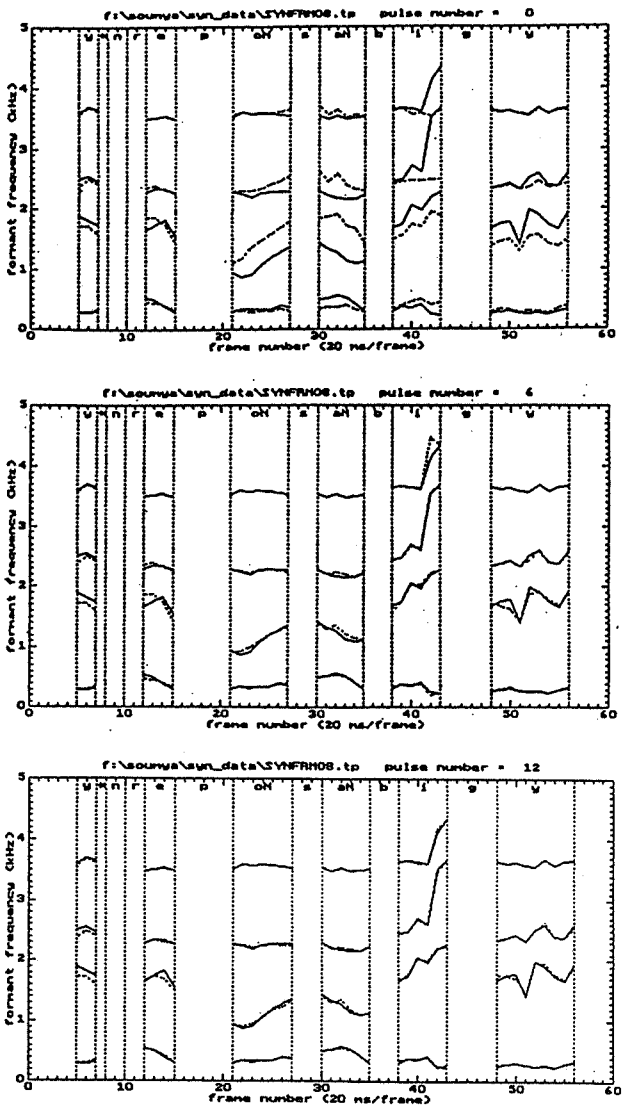


Figure 6. *Patron de formants observé (trait plein) et synthétisé en variant le paramètre* **tp** *(pointillé). a- Nombre d'impulsions : $m = 0$. b- $m = 6$. c- $m = 12$*

La distance entre la fréquence des formants observés et celle des formants calculés diminue lorsque le nombre d'impulsions augmente d'une manière analogue à celle du mouvement des paramètres articulatoires de la langue (comparer

les figures 4 et 6. Pour $m = 0$, le paramètre **tp** synthétisé est identiquement nul. Il correspond à la valeur moyenne observé). Ce premier résultat nous incite à penser que l'effet produit par les commandes des articulateurs produisent un effet direct sur les caractéristiques acoustiques.
Nous avons ensuite observé l'influence des paramètres articulatoires sur les traits acoustiques. La figure 6-a ($m = 0$ donc **tp=0** pour toute la phrase) montre l'effet de la position du corps de la langue sur l'antériorité-postériorité des voyelles. Il apparaît clairement que la distance entre les deux formants $F1$ et $F2$ est fortement biaisée lorsque le paramètre **tp** est inhibé. Nous notons ainsi une diminution de l'écart $F1 - F2$ pour les voyelles [i] et [y] et une augmentation de cet écart pour [aN] et [oN]. Ceci signifie que lorsque $\mathbf{tp} = \mathbf{0}$, les voyelles ont tendance à se neutraliser. Nous avons également constaté une relation entre la forme de la langue, exprimée par le paramètre **ts** et la protrusion des lèvres. Lorsque ce paramètre est nul, les voyelles arrondies ont un $F3$ relativement faible tandis que les voyelles non arondies voient leur $F3$ croître. Notons que le degré d'aperture (exprimé par $F1$) par rapport à la variation de la position de la mâchoire **jw** demeure inchangé. Ce résultat s'explique du fait que la locutrice PB utilise très peu sa mâchoire quand elle parle. Enfin, nous avons relevé la neutralité du paramètre **tt** sur les voyelles.

6. Conclusion

Dans ce travail, nous avons modélisé l'évolution des paramètres articulatoires de la langue par de simples filtres LPC excités par un nombre fini d'impulsions. La détection et l'estimation des impulsions sont basées sur une procédure de l'analyse-par-synthèse. L'interprétation phonétique des résultats obtenus suggère que l'organisation articulatoire au niveau neuro-musculaire entraîne la production d'un mouvement articulatoire élémentaire dont la durée est comparable à celle de la syllabe. D'autre part, la projection de cette synthèse dans le domaine formantique laisse voir une relation directe entre les variations articulatoires et les conséquences acoustiques.

Ce travail entre dans le cadre du projet européen *Esprit/BRA, n° 6975, SPEECH MAPS.*

Références

[1] Atal B.S. et Remde R., High-quality speech at low bit rates : Multi-pulse and stochastically excited linear predictive coders. Proc. ICASSP, (1986) pp. 614-617.

[2] Klatt D.H., Review of text-to-speech conversion for English. JASA, 82(3) (1987) pp. 737-793.

[3] Moulines E., Charpentier F., Pitch - synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Comm., 9 (1990) pp.453-467

[4] Fujimura O., Phonology and phonetics : A syllable-based model of articulatory organization. JASJ, (English Series) 13(1) (1992) pp. 39-48.

[5] Laferrière P., Chollet G., Miclet L. et Tubach J.P., Segmentation d'une base de données de "polysons", application à la synthèse de la parole. *XIV*$^{\text{ème}}$ JEP, 6 (1985) pp. 107-110.

[6] Maeda, S. Modélisation articulatoire du conduit vocal. Journal de Physique IV(2) (1992) pp.307-314.

[7] Bothorel A., Simon P., Wioland F. et Zerling J.P., Cinéradiographie des voyelles et consonnes du Français. Travaux de l'Institut de Phonétique, (1986).

[8] Makhoul J., Linear prediction : A tutorial review. Proc. IEEE, 63(4) (1975).

[9] Browman C.P. et McGowan R.S., Extracting dynamic parameters from speech movement data. JASA, 93(3) (1992) pp. 1580-1588.

[10] Sonoda Y. et Kiritani S., Analysis of tongue point movements by linear second-order system model. Ann. Bull. RILP, 10, (1976) pp. 29-35.

[11] Yasuhara M., Identification and Decomposition of fast handwriting system. IEEE Transaction on Circuits and Systems, 30(11) (1983) pp. 828-832.

[12] Maeda S., A digital simulation method of the vocal tract system. Speech Communication 1 (1982) pp. 199-229.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O5

PHONETIQUE

LES FACTEURS RESPONSABLES DE L'ASSIMILATION : ANALYSE DE TROIS TYPES DE GROUPES DE CONSONNES ET DE LEUR MIROIR.

CHRISTINE MEUNIER

LABORATOIRE "PAROLE ET LANGAGE" URA 261 CNRS, UNIVERSITE de PROVENCE
29, av. R. SCHUMAN, 13621 AIX-EN-PROVENCE Cedex 1

Résumé

Les phénomènes d'assimilation dans les groupes de consonnes sont complexes. Afin d'en saisir le fonctionnement, nous comparons trois groupes de consonnes et leur "miroir". Cette comparaison nous permet d'isoler les facteurs responsables de l'assimilation: la nature des unités phonétiques, le sens de l'assimilation (progressif ou régressif), le poids du facteur d'assimilation, etc. Une analyse acoustique appropriée nous permet d'identifier les phénomènes d'assimilation sous forme de segments et de sous-segments. Les résultats montrent un fonctionnement différent de la variabilité acoustique au sein des groupes de consonnes suivant qu'ils comprennent ou non une consonne vocalique.

Mots-clés: *groupes de consonnes, assimilation, variation, segments, "miroir".*

INTRODUCTION

Le groupe de consonnes est une unité à l'intérieur de laquelle les phénomènes d'assimilation sont complexes et parfois difficiles à interpréter. On évoque souvent le dévoisement des consonnes vocaliques en contexte non voisé. Ces observations sont généralement issues d'une analyse des groupes prévocaliques dans le cadre de la syllabe (autrement dit des groupes C_1C_2V où C_1 est une fricative ou une occlusive non voisée et C_2 une consonne vocalique). Il s'agit donc, dans ce cas, d'une assimilation progressive de dévoisement. Toutefois, ce type de groupe et de contexte est plus fréquement analysé dans la mesure où ce sont des groupes généralement fréquents en français [1] [2]. Dans le cadre d'une analyse de la variabilité acoustique de l'ensemble des groupes de consonnes [3], nous nous sommes interrogés sur l'origine de ce type d'assimilation: est-elle due au sens de l'assimilation (l'assimilation progressive serait plus fréquente)? Est-elle due à la nature des unités phonétiques constituantes du groupe (les consonnes vocaliques seraient moins résistantes à l'assimilation)? Est-elle due à la nature des facteurs d'assimilation (le trait non voisé aurait un "pouvoir" assimilateur plus grand)?

Afin de répondre partiellement à ces questions, nous avons traité certains groupes en position prévocalique, et leur miroir (c'est à dire leur inverse symétrique) en position postvocalique (on compare, par exemple, "plat" à "Alpes"). D'autre part, dans un corpus spécifique, et afin d'évaluer le "poids" des unités phonétiques constituantes du groupe, nous avons intégré des groupes sans consonne vocalique (groupes d'occlusive et de fricative et leur miroir). Cette analyse nous permet de comprendre plus précisément le fonctionnement des facteurs qui conditionnent l'assimilation dans les groupes de consonnes

LES CONSONNES ET LEURS COMBINAISONS

Nous classons les consonnes en trois catégories distinctes:

- "O", les occlusives [p t k b d g]
- "F", les fricatives [f s ʃ v z ʒ]
- "V", les consonnes vocaliques [l r m n j y w]

Ce classement est issu du mode d'articulation qui permet de regrouper les consonnes en fonction de leur homogénéité acoustique. Certes, il est clair que la classe des consonnes vocaliques comporte des éléments acoustiquement assez différents. Toutefois, notre objectif est de proposer un nombre de classes réduit afin de faciliter l'analyse. Or, les consonnes vocaliques ont suffisamment de caractéristiques communes (présence de structure formantique) pour que nous les regroupions dans une même classe. D'autre part, nous le verrons, ces consonnes ont un comportement similaire au contact des autres consonnes.

Ces trois classes de consonnes nous permettent de déduire, par combinaison logique, neuf classes de groupes de consonnes.

Le tableau 1 est une présentation de ces neuf classes. Parmi ces classes, certaines sont constituées de deux consonnes appartenant à la même classe de consonnes (OO, FF et VV). Nous ne nous y intéresserons pas dans cette étude.

→	Occlusive	Fricative	Cons.Voc.
Occlusive	OO	**OF**	**OV**
Fricative	**FO**	FF	**FV**
Cons.Voc.	**VO**	**VF**	VV

Tableau 1: les classes de groupes de consonnes issus de la combinaisons des trois classes de consonnes.

Les six autres groupes sont constitués de trois types de combinaison et de leur miroir (ces groupes sont en gras dans le tableau):

OV / VO
FV / VF
OF / FO

Les variations acoustiques au sein de ces six groupes seront l'objet de notre analyse dans cette communication. Seul les groupes présentant une opposition de mode de phonation (une consonne est voisée, l'autre non voisée) seront analysés.

MATERIAU LINGUISTIQUE

Pour mener à bien cette analyse, nous utiliserons deux corpus différents. Le corpus de la Base de Données des Sons du français (BDSONS [4]) comprend 5 corpus de 41 mots chacun. Ces mots sont, en quasi totalité, des monosyllabes contenant des groupes de consonnes à l'initiale (position explosive) ou en finale (position implosive). Ces mots isolés sont lus par 12 locuteurs. Ce corpus, de par sa taille, permet une analyse statistiquement importante.

Toutefois, les groupes de consonnes (ici lexicaux), ne permettent l'analyse en miroir que des groupes OV et VO. Aussi, un deuxième corpus, moins important par sa taille, nous permet d'analyser l'ensemble des six groupes énoncés plus haut. Ce corpus, appelé "Joncture" est constitué de phrases ayant toutes la même structure syntaxique: SN + SV. Les groupes de consonnes se situent à la jonction entre deux mots. Nous considérons deux niveaux de frontière:

- *Frontière majeure*: elle se situe entre le SN et le verbe. "l'équipe ralentit son allure", OV [pr]

- *Frontière mineure*: elle se situe à l'intérieur du SV, entre le verbe et le groupe nominal complément.
"ce retard handicape Robespierre", OV [pr]

Ce corpus permet de traiter toutes les combinaisons théoriques de consonnes possibles à deux niveaux de frontières syntaxiques. Les phrases sont lues par 2 locuteurs en deux répétitions. Notre analyse portera sur 84 occurrences de ce corpus.

Le corpus BDSONS nous permet de dégager les tendances observées en mots isolés pour le groupe OV et son miroir. Le corpus "Joncture" est une vérification des résultats observés sur BDSONS, et un élargissement de l'observation sur les groupes FV, OF et leur miroir.

ANALYSE ACOUSTIQUE

Afin d'analyser les phénomènes d'assimilation, nous déterminons les indices pertinents. Nous caractérisons les indices liés à trois dimensions spécifiques des consonnes:

i/ le mode d'articulation
indices: silence (occlusives), bruit (fricatives), structure formantique (consonnes vocalique)s
ii/ le mode de phonation
indices: présence ou absence de voisement.
iii/ le lieu d'articulation
indices: hauteur ou répartition des zones d'intensité renforcée sur le spectre.

La pratique de la segmentation et notre réflexion sur cette pratique [3] nous a conduit à distinguer deux types d'indices. Il est des indices qui se manifestent de façon absolue (présence ou absence); ce sont les indices de mode d'articulation et de mode de phonation. En revanche, les indices de lieu d'articulation se manifestent de façon relative et graduelle (plus ou moins).

Ces deux types d'indices distincts imposent deux méthodologies d'analyse différentes. Nous n'aborderons pas ici les indices de lieu d'articulation qui ont été l'objet de travaux précédents [5] et [6].

Le caractère absolu des indices de mode d'articulation permet d'envisager la variation, ainsi que les phénomènes d'assimilation, sous forme de segments (unités phonétiques) et de sous-segments (segments acoustiquement homogènes). Si l'on considère que chaque classe de consonnes définie plus haut est caractérisée par des définitions canoniques, toute non-réalisation de ces formes sera considérée comme variation.

Les trois classes de consonnes sont définies selon le mode d'articulation et possède donc les indices carcatéristiques énoncés plus haut (voir, "i/ le mode d'articulation"). Les occlusives et fricatives sont également caractérisées par leur voisement ou leur non voisement. Ainsi, la description canonique de la consonne [f] sera: *présence de bruit + non voisement.*

Toute inversion de ces caractéristiques sera donc une variation ou une assimilation.

Description canonique	**Indices de variation**
consonanticité (silence bruit)	*vocalisation* (apparition de structure formantique)
vocalicité (structure de formant)	*consonantisation* (disparition de la structure formantique apparition de bruit)
voisement (F0>0)	*dévoisement* (F0=0)
non voisement (F0=0)	*voisement* (F0>0)

Tableau 2: représentation des différentes formes canoniques et des indices de variation. Ces indices se présentent comme des inversions des formes canoniques.

Les phénomènes d'assimilation seront donc identifiés comme des segments ou des sous-segments (que nous appelerons "segments transitoires") localisés à la frontière entre deux unités phonétiques (voir les figures 7 et 8 qui sont des exemples de segments transitoires, à la fin de cette communication). Ils seront attribués à l'une ou l'autre consonne selon des critères de segmentation [3]

RESULTATS DE L'ANALYSE

Corpus BDSONS

La présence de segments transitoires n'est pas due au mode d'articulation des consonnes mais plutôt au mode de phonation des occlusives ou des fricatives (figure 1). Les consonnes vocaliques se comportent sensiblement de la même façon au contact des occlusives et des fricatives. Le dévoisement est clairement le phénomène de variation le plus fréquent dans les groupes où les consonnes vocaliques sont en contexte non voisé. Le segment transitoire n'affecte que les consonnes vocaliques.

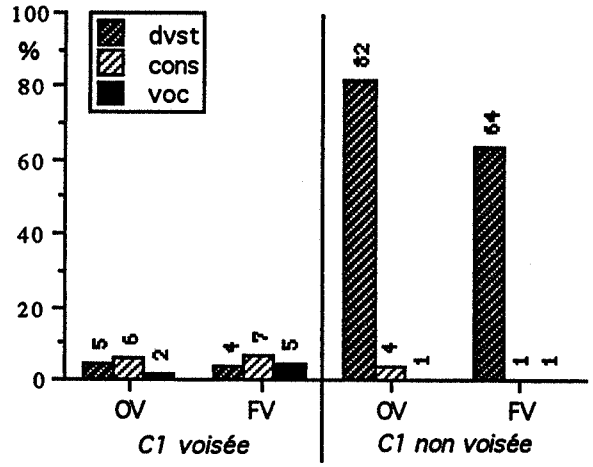


Figure 1: pourcentage de chaque type de segments transitoires dans les deux groupes OV et FV en fonction du voisement de l'occlusive et de la fricative.

Lorsque l'on compare les groupes miroirs, on constate que le nombre de segments transitoires est plus important dans les groupes OV que dans les groupes VO (figure 2). Une observation précise de la proportion des segments transitoires lorsque l'occlusive est non voisée permet de dégager des tendances différentes à l'intérieur des deux groupes. Dans le groupe OV, le dévoisement de la consonne vocalique est très fréquent alors que la consonantisation est très rare. A l'inverse, la proportion de dévoisement dans le groupe VO est bien moins fréquente tandis que le nombre de consonantisation augmente (figure 2).

Ainsi, l'assimilation progressive de dévoisement (dans les groupes explosifs) est plus systématique que l'assimilation régressive (dans les groupes implosifs). Le fonctionnement est inverse pour le phénomène de consonantisation. Dans les deux groupes, la consonne vocalique est la seule à être affectée.

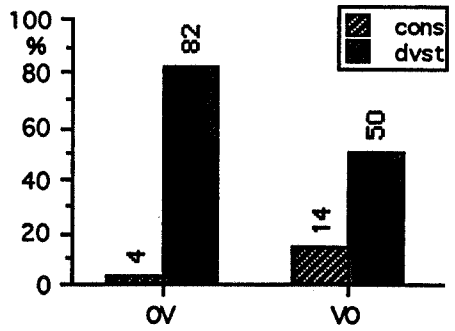


Figure 2: pourcentage de chaque type de segments transitoires dans les deux groupes OV et VO lorsque l'occlusive est non voisée.

Corpus "Jonction"

En raison de la différence des phénomènes d'assimilation, nous distinguerons les groupes comprenant une consonne vocalique des groupes n'en comprenant pas.

L'analyse des groupes OV, FV et de leur miroir confirme les résultats observés au cours de l'analyse du corpus BDSONS. L'assimilation progressive de dévoisement est systématique dans les groupes où la consonne vocalique est en position C2. En revanche, lorsque cette consonne est en position C1, l'assimilation est moins fréquente et se manifeste sous forme de consonantisation (figure 3)

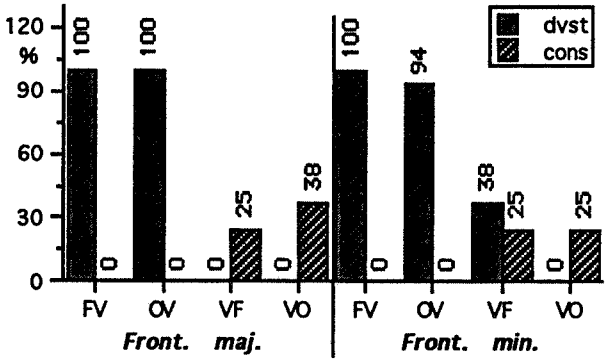


Figure 3: Proportion du type de segments transitoires (dévoisement et consonantisation) pour les deux groupes FV et OV et leur "miroir" dans les deux types de frontière lorsque l'occlusive ou la fricative sont non voisées.

Il semble donc que ces phénomènes soient réguliers: on les retrouve aussi bien en mots isolés qu'en jonction de mots et quel que soit le type de frontière entre les deux consonnes. Les facteurs d'assimilation dans les groupes où il y a consonne vocalique sont donc:

- le mode de phonation des consonnes (le non voisement des fricatives et des occlusives est responsable de la présence des segments transitoires, figure 1).
- le mode d'articulation des consonnes (seules les consonnes vocaliques sont affectées).
- le sens de l'assimilation (la consonne vocalique est affectée de façon plus importante lorsqu'elle est en position C2, figures 2 et 3)

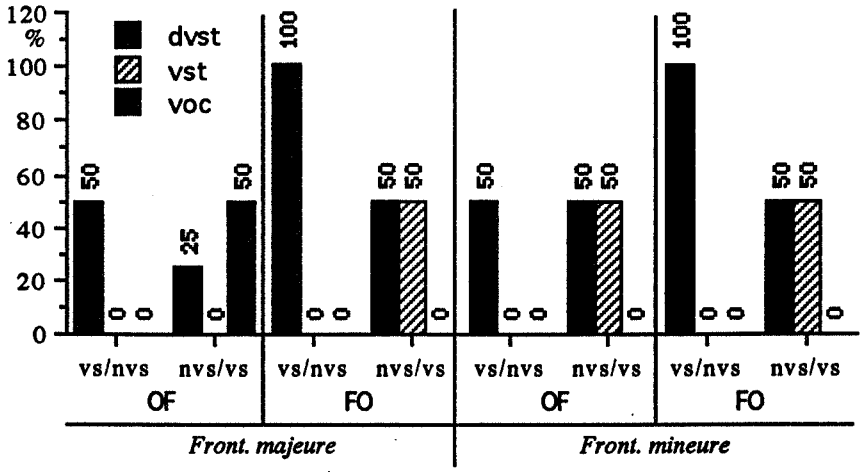


Figure 4: Proportion du type de segments transitoires (dévoisement et consonantisation et vocalisation) pour le groupe OF et son "miroir" dans les deux types de frontière.

Les groupes de consonnes ne comprenant pas de consonne vocalique se comportent de façon différente.

Lorsqu'il y a opposition de voisement dans le groupe, la proportion de segments transitoires est variable. Il s'agit exclusivement de dévoisements lorsque la consonne voisée est en position C1, quelle qu'elle soit (occlusive ou fricative). Lorsque C1 est non voisée, on constate des assimilations de voisement et de dévoisement (figure 4).

Ainsi, l'assimilation regressive de dévoisement est régulière, tandis que dans les groupes non voisé + voisé, l'assimilation de mode de phonation est soit regressive, soit progressive. Les deux positions, C1 comme C2, sont affectées. Les deux types de consonnes, les occlusives comme les fricatives (dont le dévoisement est plus systématique en position C1), sont également affectées. On notera que, dans les groupes non voisé + voisé, l'un des deux locuteurs "dévoise" alors que l'autre "revoise". La stratégie des locuteurs semble donc jouer un rôle concernant le sens de l'assimilation.

Les facteurs d'assimilation dans les groupes où il n'y a pas de consonnes vocaliques sont:

- le mode de phonation (dans les groupes où il n'y a pas opposition de mode de phonation, les segments transitoires sont quasi inexistants).
- le sens de l'assimilation lié à la nature des facteurs d'assimilation (le dévoisement est plus systématique lorsqu'il est régressif).
- la stratégie du locuteur (le voisement ou dévoisement peut éventuellement être fonction du locuteur).

On notera que le mode d'articulation des deux consonnes n'est pas un facteur de variabilité pertinent.

Les figures 5 et 6 montrent que les consonnes vocaliques sont plutôt affectées en position C2 (donc de façon progressive) alors que les occlusives et fricatives sont plutôt affectées en position C1 (donc de façon régressive), quel que soit le type de frontière.

Il semble donc que le sens de l'assimilation soit fonction de la nature des unités phonétiques en contact.

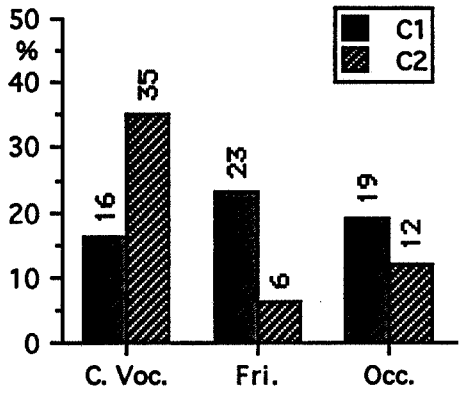


Figure 5: Proportion de segments transitoires pour les trois classes de consonnes en position C1 ou C2 en frontière majeure.

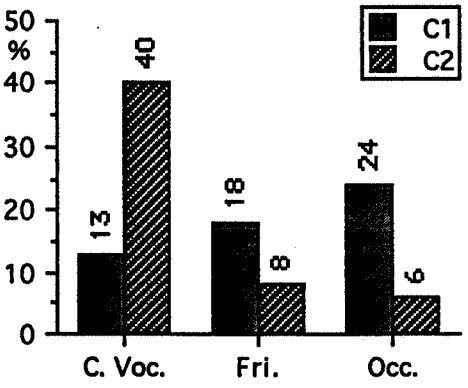


Figure 6: Proportion de segments transitoires pour les trois classes de consonnes en position C1 ou C2 en frontière mineure.

On observe donc un fonctionnement différent des phénomènes d'assimilation suivant que les groupes de consonnes comprennent ou ne comprennent pas de consonne vocalique.

DISCUSSION

Les résultats de cette analyse montrent que l'assimilation de dévoisement des consonnes vocaliques en contexte non voisé est due à deux types de facteurs:

- leur nature phonétique: elles sont les seules à être affectées systématiquement (alors que dans les groupes OF et FO les occlusives et les fricatives sont affectées indifféremment). Nous devons préciser toutefois que le comportement des consonnes vocaliques n'est pas homogène: la consonne /r/ est celle qui subit le plus fortement les phénomènes assimilatoires (dévoisement et consonantisation).
- le sens de l'assimilation (elles sont plus fortement affectée en position C2 qu'en position C1, l'assimilation est donc plus importante lorsqu'elle est progressive).

Si l'on se place dans une optique de variabilité acoustique, ces observations confirment une division des groupes de consonnes en deux classes primaires: les groupes avec consonne vocalique et les groupes sans consonne vocalique. Ce phénomène est constaté au travers d'autre analyses [3]. La variabilité au sein des groupes de consonnes fonctionne différemment suivant qu'une consonne vocalique est, ou n'est pas, présente dans le groupe.

- présence d'une consonne vocalique: les phénomènes de variation sont réguliers. Les consonnes vocaliques subissent systématiquement l'influence du contexte (on peut penser ici à la "force" des sourdes et à la faiblesse des "sonores" [7]). Les consonnes vocaliques fonctionnent comme un élément régulateur, dans le sens où elles "assument" régulièrement les source de variabilité.

- absence d'une consonne vocalique: les phénomènes de variation sont moins réguliers. Ils ne dépendent pas de la nature des unités phonétiques, mais d'autres facteurs: le sens et la nature du facteur de variabilité (l'assimilation régressive de dévoisement est plus fréquente [8]); la stratégie des locuteur (au sein de ces groupes, on constate plus des tendances que des phénomènes réguliers; tout type d'assimilation est possible, progressive ou régressive, de voisement ou de dévoisement [9]). Ces observations signifient que les occlusives et les fricatives résistent pareillement au phénomènes d'assimilation. Aussi, ces phénomènes sont-ils moins réguliers et dépendent-ils de facteurs externes à la nature des consonnes du groupes. Ils n'y a donc pas d'éléments régulateurs.

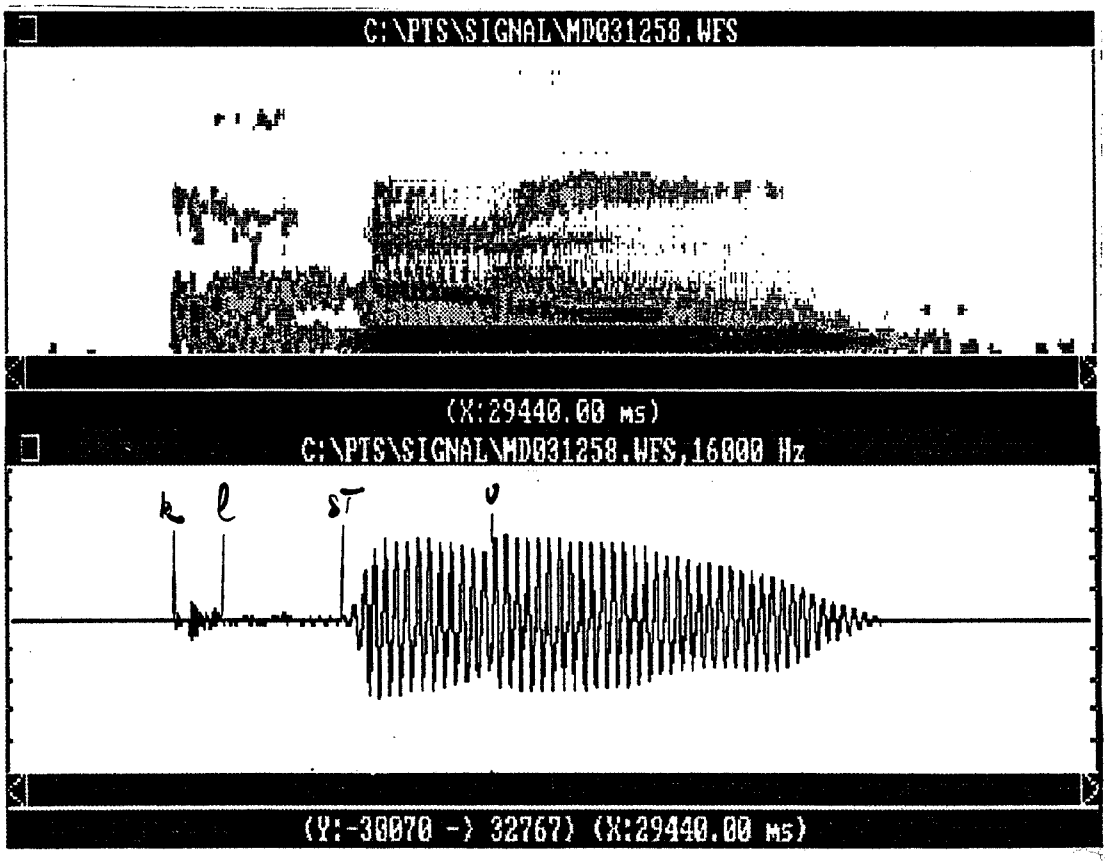


Figure 7: /klu/ prononcé par le locuteur MD de la BDSONS. Le /l/ est réalisé partiellement dévoisé.

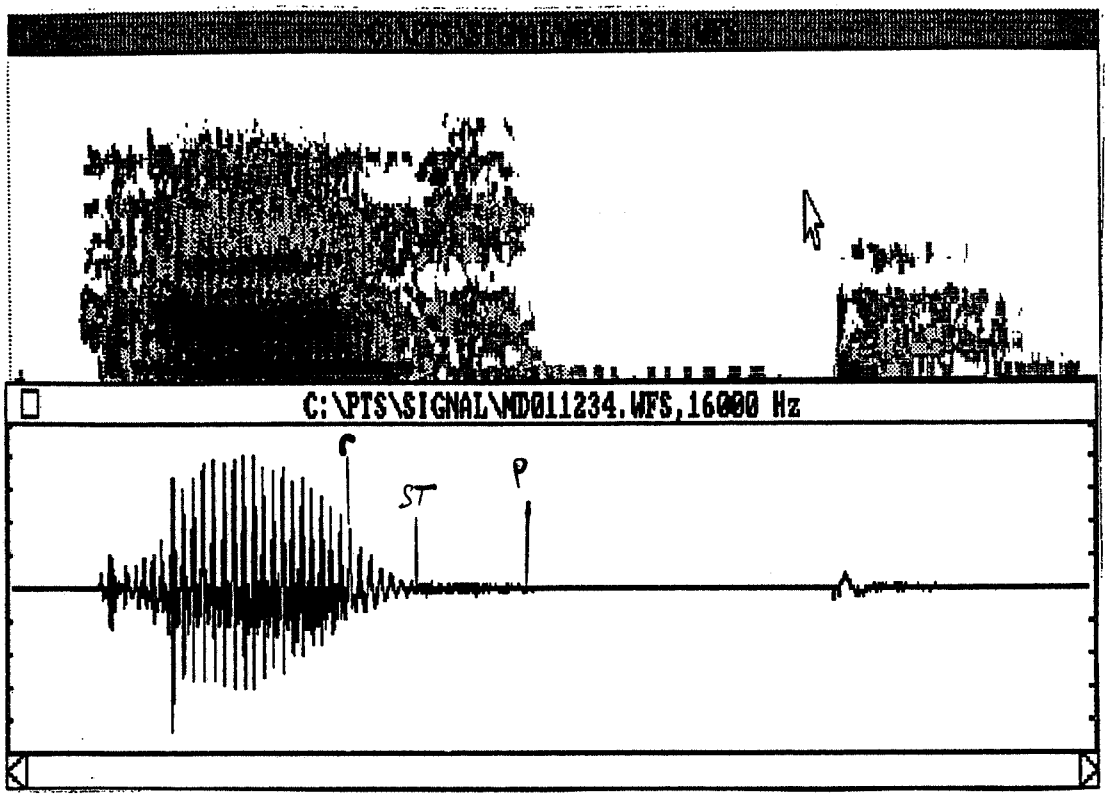


Figure 8: /arp/ prononcé par le locuteur MD de la BDSONS. Le /r/ est réalisé partiellement dévoisé.

CONCLUSION

L'analyse que nous avons présenté ici tend à montrer que les phénomènes de variation au sein des groupes de consonnes ne fonctionnent pas de façon homogène. Autrement dit, il n'y a pas de règles générales de variation commune à l'ensemble des groupes de consonnes. Ces phénomènes de variation ne dépendent pas non plus uniquement de la nature des unités phonétiques en contact (on constate les mêmes phénomènes pour OV et FV, d'une part, et pour VO et VF, d'autre part). En fait, il est possible d'attribuer les variations à une conception intermédiaire: les groupes de deux consonnes ne fonctionnent pas comme une unité acoustique homogène; ils ne fonctionnent pas non plus comme une simple concaténation d'unités; en revanche, les groupes se divisent en deux unités acoustiques à l'intérieur desquelles les phénomènes de variation ont un fonctionnement propre.

REFERENCES BIBLIOGRAPHIQUES

[1] ROSSI, M., 1968, "Au sujet des groupes consonantiques du français", *Revue d'Acoustique*, 3, 4, pp. 306-311.

[2] AUBERGE, V., BOË, L.J., LEFEVRE, J.P., 1988,"Lexique et groupes consonantiques", *Actes des 17èmes Journées d'Etude sur la Parole*, Nancy, pp. 55-60.

[3] MEUNIER, C. ,1994, *Les groupes de consonnes : problématique de la segmentation et variabilité acoustique*, Thèse de l'université de Provence, Présenté le 7 mars 1994.

[4] CERVANTES, O., SERIGNAT, J.F., DESCOUT, R., CARRE, R., 1986, "Définition et réalisation d'une base de données des sons du français", *Actes des 15èmes Journées d'Etude sur la Parole*, Aix-en-Provence, pp. 213-216.

[5] MEUNIER, C. ,1992, "Variabilité acoustique des groupes de consonnes: le rôle de la voyelle adjacente"*Actes du 12ème Congrès Français d'Acoustique*, Arcachon, pp. 299-302.

[6] MEUNIER, C. ,1992, "Importance des différents facteurs de variabilité interne au groupe de consonnes", *Actes des 19èmes Journées d'Etude sur la Parole*, Bruxelles, pp. 475-480.

[7] SIMON, P., 1967, *Les consonnes françaises. Mouvements et positions articulatoires à la lumière de la radiocinématographie*, Paris, Klincksieck (eds).

[8] MARCHAL, A., DEL NEGRO, A.S., 1991, "Gémination phonétique en frontière de mots", *Proceedings of the 12th International Congress of Phonetic Sciences* , Aix-en-Provence, vol. 4, pp. 438-441.

[9] ROCHETTE, C., 1973, *Les Groupes de consonnes du français: étude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie*, Klincksieck, Québec-Paris.

CORRELATS ACOUSTIQUES ET PERCEPTION DE LA DIPHTONGAISON DES VOYELLES FRANÇAISES /e/ ET /o/ PAR LES LOCUTEURS AMERICAINS.

Bertrand LAURET

Institut de Phonétique, CNRS URA 1027
Université de la Sorbonne Nouvelle, Paris III
19, rue des Bernardins
75005-Paris.

Résumé

L'objectif de cette étude est d'apporter des données de phonétique acoustique permettant d'observer la diphtongaison (variation du timbre) au cours de l'émission des voyelles orales françaises /e/ et /o/ par les locuteurs américains. Des enregistrements d'énoncés français lus par deux locuteurs américains et deux locuteurs français ont permis de comparer des séquences en français de V1cV2# (V1=V2=/e, o/ et C=/p, b, s, z/). Ces séquences étaient présentées dans différents contextes intonatifs. Un jury de 5 phonéticiens a évalué la présence de diphtongaison et/ou de variation de la fréquence fondamentale des réalisations des locuteurs américains en français: 43,3% des voyelles (V1) devant consonne et 93,7% des voyelles finales (V2) ont été perçues par le jury comme étant diphtonguées.
Une analyse formantique comparant les réalisations diphtonguées et non-diphtonguées des locuteurs américains et les réalisations des locuteurs français a montré que les critères performants pour distinguer, dans ce corpus, les réalisations perçues comme étant diphtonguées sont: l'amplitude de la variation de F2 au cours de la voyelle pour /e/ (supérieure à 400Hz, la voyelle est perçue comme étant diphtonguée) et l'amplitude de la variation du centre de gravité CG (CG=(F1+F2)/2) au cours de la voyelle pour /o/ (supérieure à 100-120Hz, la voyelle est perçue comme étant diphtonguée) Aucune relation entre la perception de la diphtongaison et les variations de la fréquence fondamentale n'a pu être établie.

Mots-clés: diphtongaison, anglais-français.

INTRODUCTION

La réalisation diphtonguée des voyelles françaises /e/ et /o/ par les locuteurs américains est un important facteur porteur de l'accent américain en français (Le Bras, 1981; Tranel, 1987). Comme illustration préliminaire du type d'effet étudié ici, on peut comparer *c'estchaud* et *say show* (Figure 1).

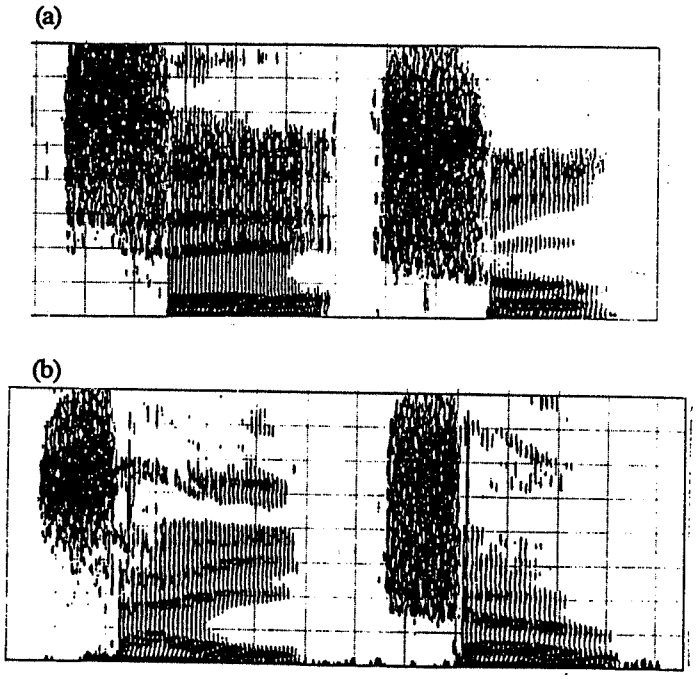


Figure 1. Sonagrammes des phrases *c'estchaud* par un locuteur français (a) et *say show* par un locuteur américain (b).

L'hypothèse générale suivie par l'ensemble des études sur "l'accent étranger" est que la structure phonologique et les caractéristiques phonétiques de la langue maternelle (L1) vont influencer la prononciation des sons d'une langue étrangère (L2). Le terme "accent étranger" recouvre alors toutes les caractéristiques d'une prononciation qui diverge du standard général de la langue dans la production d'une personne n'appartenant pas à la communauté linguistique concernée (Léon, 1971; Kassaï, à paraître).
Troubetzkoy (1939) est apparemment le premier à faire l'hypothèse que la phonologie de L1 cause un filtre dans l'apprentissage de L2, en particulier pour les différences perceptives qui ne sont pas phonologiques en L1.

Suivant la perspective contrastive de Lado (1957), un apprenant L2 va transférer les phonèmes L1 et leurs variantes, les schémas rythmiques et les schémas intonatifs de L1.
Flege (1984) propose un traitement différent des phonèmes de L2 nouveaux (n'ayant pas de correspondant en L1) et des phonèmes de L2 similaires (ayant un correspondant en L1). Il prédit par une analyse contrastive suivant les principes phonologiques que, si le nouveau son de L2 est composé de traits qui spécifient des sons de L1, ce son sera appris sans difficultés.
Munro (1993) cherchant à mettre en relation les mesures acoustiques effectuées sur des items anglais prononcés par des arabophones et les jugements de bonne prononciation ("native-like") rappelle la diversité d'objectifs et de méthodologies des études sur l'accent étranger.

1. LES VOYELLES /e/ ET /o/ EN ANGLAIS ET EN FRANÇAIS.

1.1. STATUT PHONOLOGIQUE

Les voyelles de l'anglais sont généralement caractérisées, au contraire des voyelles du français, par l'instabilité de leur timbre. Si, théoriquement, toutes les voyelles anglaises sont susceptibles d'être diphtonguées (en particulier après un <r> final réalisé en diphtongue décroissante avec un deuxième élément /ə/ ou /ə˞/, Heffner, 1964), certaines ont élevé la diphtongaison au rang de trait phonologique distinctif.
Les voyelles /e/ et /o/ sont unanimement reconnues comme particulièrement sensibles à la diphtongaison en particulier en position syllabique accentuée. Qu'elles soient étudiées avec le plein statut de diphtongue (par exemple, en tant que /eɪ, oʊ/, Ladefoged, 1975) ou comme voyelles diphtonguées (Gimson, 1962), les voyelles anglaises /e/ et /o/ sont décrites comme particulièrement sujettes à variation entre différents types régionaux et sociaux de réalisations. Peterson et Coxe (1953), dans leur article consacré à ces deux voyelles, citent les "Principes de l'Association Internationale de Phonétique" (1949) qui précisent que les variétés britanniques et américaines de l'anglais diphtonguent relativement le /e/ et le /o/.

1.2. DONNEES ACOUSTIQUES

Les diphtongues sont généralement définies comme un noyau syllabique présentant, indépendamment des phones précédants ou suivants, "un mouvement d'une voyelle à une autre" (Ladefoged, 1975), ou comme "un élément syllabique commençant par un son et glissant vers un autre" (Heffner, 1964). Alors que les formants de syllabes contenant des voyelles "simples" ou "pures" passent par la position d'une voyelle particulière, les formants des syllabes contenant des diphtongues passent généralement par au moins deux positions successives (Peterson, 1953). Lorsqu'une voyelle non diphtonguée se trouve dans un contexte consonantique, les transitions de formants peuvent être expliquées par la cible de cette voyelle et les mouvements consonne à consonne qui se superposent et éventuellement par les phénomènes de coarticulation entre voyelles: on peut expliquer les mouvements de formants en fonction de la cible de la voyelle et de son contexte. Par contre, pour les diphtongues, les mouvements des formants durant la partie vocalique ne peuvent pas être seulement expliqués par la superposition de mouvements consonne à consonne et d'une seule cible vocalique (ou direction vers une cible) sous-jacente mais par un glissement d'une cible à l'autre.
La nature dynamique des diphtongues a été depuis longtemps reconnue au travers des pratiques de transcription en utilisant des symboles double. Dans une langue très étudiée comme l'anglais, l'absence de consensus pour transcrire les diphtongues démontre la complexité du problème, et illustre le besoin d'une plus grande précision des données acoustiques.
En fait, contrairement aux transcriptions traditionnelles, les diphtongues ne commencent ni ne finissent avec aucune des voyelles simples de la langue (Gay, 1968; Ladefoged, 1975). Pour Delattre (1951), si une voyelle est telle que son point de départ et son point d'aboutissement rappellent assez nettement les timbres de deux voyelles différentes de la langue étudiée sans qu'il y ait pour autant deux syllabes, on dit alors que c'est une diphtongue. Pour Lehiste (1967), s'il y a deux cibles vocaliques entre les consonnes finale et initiale dans des combinaisons du type C-Noyau Syllabique-C, le noyau syllabique est une diphtongue. De plus, l'une des deux parties doit être nettement plus accentuée que l'autre, car la diphtongue entière forme le nucleus inséparable d'une seule syllabe (Delattre, 1965). Seul Fry (1979) apporte la contradiction à la notion de cible non-atteinte en définissant la diphtongue comme la juxtaposition de deux voyelles simples. Pour lui, et pour lui seul, les spectrogrammes des diphtongues montrent clairement et successivement les caractéristiques des deux voyelles.
Les changements de couleur sonore dans ce type de voyelles sont principalement produits par le mouvement continu des articulateurs de la parole. Selon la théorie acoustique de la production, à une matrice de valeurs formantiques données correspond un ensemble de formes du conduit vocal. Les caractéristiques de la variation temporelle des valeurs formantiques représentent donc les traits dynamiques des diphtongues et plus généralement des voyelles "composées". C'est pourquoi les diphtongues sont souvent représentées par des flèches placées sur des triangles vocaliques sans valeurs-cible des fréquences des formants, alors que la notion de cible des voyelles pures permet d'une part de représenter ces voyelles comme des points sur les triangles vocaliques, et de donner les valeurs des fréquences des formants.
Hertz (1991), reprenant les expériences de Gay (1968) en synthèse de la parole, distingue les diphtongues des "gliding vowels" . Les diphtongues /ɔɪ, aɪ, aʊ/ sont synthétisées, au même titre qu'une séquence VV, par deux cibles formantiques stables et par une longue transition interne. Les "gliding vowels" /eɪ, oʊ /

(voire /uu, ii/) consistent en un premier lieu en une transition: les cibles formantiques initiales et finales n'ont aucune durée.

2. EXPERIENCE

Le but de cette étude est de confirmer la difficulté qu'ont les locuteurs américains à prononcer /e/ et /o/ en français, et d'observer les caractéristiques acoustiques des réalisations vocalique des locuteurs américains. L'hypothèse est que le contexte suprasegmental peut favoriser l'apparition d'une marque vocalique d'accent anglophone.

2.1. CORPUS

Un corpus d'énoncés français a été établi en répondant aux préoccupations suivantes: des énoncés courts ne présentant la combinaison étudiée (V1CV2; V1=V2=/e, o/, et C=/p, b, s, z/) qu'une seule fois pour éviter les "tongue twisters" à la Delattre; des énoncés ayant un sens; des énoncés présentant la combinaison VCV dans quatre schémas intonatifs: finale d'énoncé (exemple: "C'est une *épée*"); enchaînement (exemple: "C'est une *épée* dorée"); montée de continuation majeure, par mise en apposition (exemple: "Cette *épée*, je l'ai cassée"); montée de continuation réduite (ou mineure) en fin de syntagme nominal (exemple: "Cette *épée* pourra te servir"). Ces différents contours intonatifs influençant la durée de la dernière syllabe, les incidences d'un allongement sur le phénomène de diphtongaison pouvaient être observées. Les mots présentés sur ces schémas intonatifs sont: *épée, bébé, fessée, lésé, au pot, bobo, au saut, zozo.*

2.2. INFORMATEURS ET ENREGISTREMENT

Deux locuteurs américains vivant en France ayant appris le français au cours de leur scolarité, d'un bon niveau de français, (une femme de 36 ans, originaire du Massachusetts, en France depuis 3 ans et un homme de 31 ans, originaire du Wisconsin, en France depuis 3 ans) et deux locuteurs français (une femme de 25 ans, originaire de Lorraine, et un homme de 28 ans, originaire de Paris) ont été enregistrés en chambre sourde avec un magnétophone Marrantz CP 130 muni d'un micro à électret. La consigne, lors de l'enregistrement des énoncés, était de "lire lentement et distinctement comme pour un exercice de prononciation".

2.3. MESURES

Les enregistrements analogiques ont été numérisés sur un PC muni du logiciel d'analyse du signal UNICE du LIMSI-VECSYS. Les mots portant les séquences V1CV2 ont été extraits. La durée des voyelles V1 et V2 a été mesurée. Les valeurs de la fréquence fondamentale et des trois premiers formants ont été relevées toutes les 10 millisecondes à partir de l'explosion des consonnes occlusives et des premières vibrations périodiques suivant les consonnes fricatives. Ces mesures ont été effectuées par détection automatique de la fréquence fondamentale, et par analyse LPC des spectres instantanés pour les mesures des fréquences des formants.
Dans le cadre d'une étude contrastive des "gestes" vocaliques, il nous a semblé utile de relever les valeurs des fréquences des formants tout au long de la voyelle. Des comparaisons entre ce type de mesures "manuelles" et des mesures effectuées par un logiciel de détection automatique développé à l'ENST dont nous disposions (Chennoukh et Carré, Analyseur de signaux, 1992) ne nous ont pas amené à utiliser la détection automatique: en effet, les nombreuses erreurs effectuées par la détection automatique, dues à la diversité des spectres instantanés et surtout à l'évolution des formants au cours de la voyelle, ne sont pas toujours éliminables. Nous avons préféré nous tenir à l'observation minutieuse et manuelle des évolutions formantiques.

2.4. EVALUATION DE L'ACCENT

Pour évaluer la diphtongaison des voyelles, autrement dit une marque d'accent anglophone en français, il ne semblait pas possible de présenter les séquences V1CV2 complètes à un jury. En effet, les consonnes retenues (occlusives bilabiales et fricatives dentales) présentant des caractéristiques différentes en anglais et en français (VOT des occlusives sourdes, voisement des occlusives sonores, lieu d'articulation des fricatives alvéolaires / dentales), elles étaient susceptibles d'être, elles aussi, porteuses de l'accent anglophone. C'est pourquoi seules les voyelles V1 et V2 extraites des séquences prononcées par les locuteurs américains ont été utilisées comme stimuli pour l'élaboration du test d'évaluation effectué auprès d'un jury de 5 étudiants phonéticiens: les voyelles V1 et V2 étaient proposées séparément. La question posée était: "Cette voyelle est-elle prononcée par un(e) Français(e)?". Lorsqu'au moins 3 membres du jury estimaient qu'une voyelle de test n'était pas produite par un locuteur français, cette voyelle était considérée comme portant une marque d'accent anglophone. 43,3% des voyelles situées devant consonne (V1) et 93,7% des voyelles finales (V2) ont ainsi été perçues par le jury comme étant produites par des locuteurs étrangers. Les V1 sont globalement perçues comme plus porteuses d'une marque d'accent étranger devant les consonnes voisées, car elles subissent alors un allongement. Mais les différents contours intonatifs n'ont pas influencés de façon significative la perception du phénomène pour les V2, ces dernières ayant été dans la plupart des cas (93,7%) perçues comme portant une marque d'accent étranger.

2.5. DUREE

Les mesures de durée des V1 et des V2 des 4 locuteurs n'ont permis de tirer aucune conclusion, la durée des voyelles dépendantde des vitesses d'élocution de chaque locuteur. Les durées vocaliques des deux hommes (américain et français) étaient très similaires, alors que les durées vocaliques de la locutrice

américaine étaient beaucoup plus importantes que celles de la locutrice française. Pour les V1, l'influence du voisement de la consonne suivante a été observée. Pour les V2, les différents contours intonatifs ont effectivement influencés la durée de la voyelle finale (Figure 2).

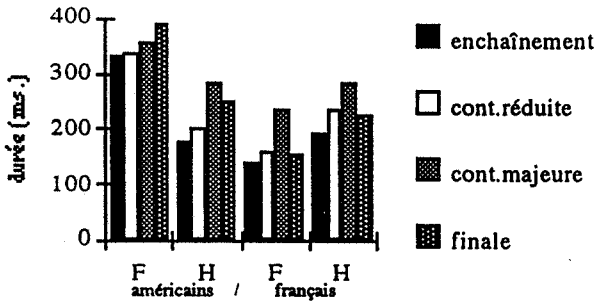


Fig. 2 - Durées moyennes en ms. des V2 dans des séquences V1CV2 où V1=V2= /e, o/ et C= /p, b, s, z/.) par locuteur et selon le contour intonatif

2.6. FORMANTS

Pour l'analyse formantique, on a comparé des voyelles des locuteurs américains, certaines ayant été jugées comme porteuses de l'accent anglophone, d'autres n'ayant pas été jugées comme porteuses de l'accent anglophone, ainsi que des voyelles des locuteurs français.

2.6.1. V1=V2=/e/, C=/p,b/

L'analyse des réalisations des locuteurs américains montre que le deuxième formant de la voyelle /e/ permet de distinguer les réalisations marquées des réalisations non-marquées. Le schéma global de variation de F2 est différent pour les locuteurs américains et pour les locuteurs français. Ces deux types de variation de F2 permettent néanmoins de dégager trois points d'observation communs: une valeur initiale qui correspond à la valeur minimale; une valeur maximale qui se trouve plutôt au-delà du milieu de la voyelle pour les réalisations des locuteurs américains et plutôt au milieu de la voyelle pour les locuteurs français; une valeur finale. On observe une plus grande symétrie du schéma général des variations de F2 des réalisations des locuteurs français. Cela se traduit par une plus grande similitude des valeurs de F2 au début et à la fin de la voyelle. Par contre, pour les réalisations des locuteurs américains, on note une grande variation de fréquence entre F2 initial et F2 maximum (d'autant plus grande que la réalisation est marquée), suivie d'une moindre amplitude de variation (négative) entre F2 maximum et F2 final (plutôt plus importante quand la réalisation est marquée).
La variation de fréquence de F2-F1 est un critère moins performant pour identifier les réalisations de notre corpus perçues comme étant diphtonguées (voir Tableau 1).

	américain marqué		américain non-marqué				français				
V1: Fq. F2	508	547	391	312	352	312	118	117	94	156	125
Fq. F2-F1	507	625	430	351	312	312	118	117	126	156	125
V2: Fq. F2	406	507			156				250		
Fq. F2-F1	400	468			234				187		

Tableau 1. Variations de fréquence de F2 (Hz) et de (F2-F1) pour les V1 et V2=/e/ produites par les locuteurs américains (réalisations marquées et non marquées) et par les locuteurs français.

Ces résultats sont résumés dans la Figure 2:

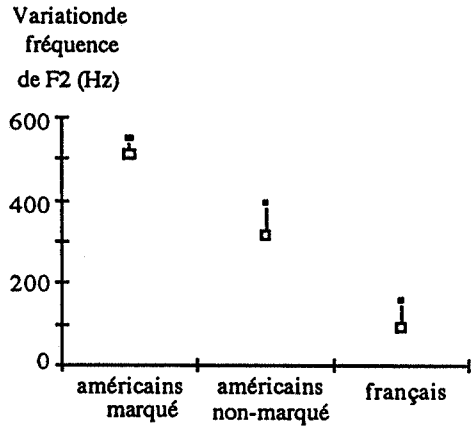


Fig. 3 - Variation de fréquence de F2 des réalisations de la voyelle /e/

2.6.2. V1=V2=/o/, C=/p,b/

Un changement de timbre de /o/ vers /u/, autrement dit une dynamique fermante de la voyelle /o/, se traduit par un abaissement de F1, une relative stabilité (plutôt descendante) de F2 et une baisse de F3. On sait que, pour des raisons de lois acoustiques, l'amplitude de F3 est faible pour les voyelles postérieures, ce qui est dû aux formants F1 et F2 qui sont bas pour ces voyelles.
La synthèse des voyelles par deux formants réalisée dans les Laboratoires Haskins a montré qu'une voyelle pouvait être produite avec F1 et un deuxième formant supérieur au formant F2 que Fant (1973) appelle F ' 2, défini de sorte à considérer d'autant plus F3 que F2 augmente en fréquence. La formule utilisée est:

$$F'2 = F2 + \frac{1}{2}(F3-F2)\frac{(F2-F1)}{(F3-F1)}$$

Quand F2 est très proche de F1, on obtient une valeur limite de F ' 2, soit F2.
Les voyelles postérieures peuvent être synthétisées par un seul formant, car elles présentent naturellement les deux premiers formants séparés de moins de 3,5 barks en fréquence: le formant unique de la voyelle ainsi synthétisée doit alors correspondre au centre de gravité

du spectre naturel restreint à la zone des deux premiers formants (Calliope, 1989). On a montré que, pour les voyelles postérieures, F1 et F2 ne seraient pas séparés du point de vue perceptif, mais qu'ils seraient inclus dans une même bande critique. A la recherche d'un paramètre conjuguant les baisses de F1 et de F2, nous définirons le centre de gravité comme:

CG = (F1+F2)/2 (Hz)

L'analyse des réalisations de /o/ a rencontré les problèmes suivants: le regroupement des deux premiers formants dans les basses fréquences rend leur identification plus difficile; la faible amplitude de F3 nous a conduit à chercher des indices de diphtongaison dans l'évolution temporelle des seuls deux premiers formants.

L'observation de l'évolution temporelle de F1 nous a permis de dégager deux schémas globaux de variation: un début de voyelle avec des valeurs de fréquence de F1 maximum, puis décroissantes pour les réalisations marquées; une symétrie par rapport au centre des réalisations non-marquées. Or, les valeurs très basses des fréquences de F1 au contact des consonnes des réalisations du locuteur français engendrent des amplitudes de variation du même ordre que celles des réalisations diphtonguées des locuteurs américains. Les amplitudes des variations fréquentielles de F1 n'ont pas permis de distinguer les réalisations marquées des réalisations non-marquées.

Si l'amplitude de F2 n'a pas été un critère performant, l'observation des amplitudes des variations du centre de gravité permet de distinguer les réalisations marquées des autres réalisations mais ne permet pas de distinguer l'ensemble des réalisations des locuteurs américains de l'ensemble des réalisations des locuteurs français (Figure 4).

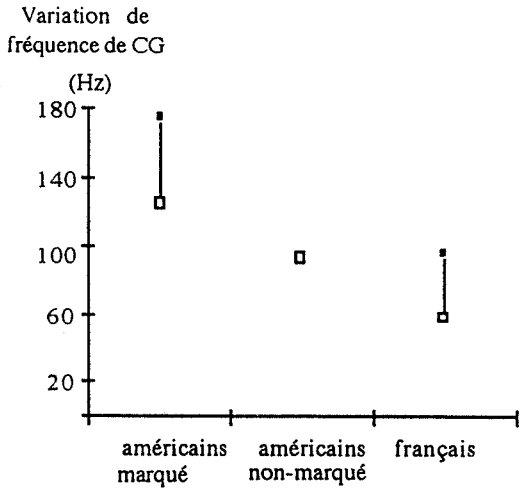


Fig. 4 - Variation de fréquence de CG = (F1+F2)/2 des réalisations de la voyelle /o/

CONCLUSION

La recherche d'indices acoustiques caractérisant la diphtongaison des voyelles orales françaises /e/ et /o/ par les locuteurs américains nous a conduit à explorer en détail la dynamique formantique d'une partie des réalisations vocaliques que nous avons enregistrées. On sait, en effet, que l'information essentielle pour l'identification des voyelles réside dans la forme globale du spectre et plus précisément dans la localisation des maxima spectraux correspondant aux deux ou trois premiers formants. L'information fréquentielle, pour la perception de la distance phonétique entre les voyelles, est plus importante que la pente globale du spectre et que les largeurs de bande des formants. Le caractère dynamique des réalisations diphtonguées exigeait donc une étude des variations temporelles des fréquences des formants.

Les caractéristiques formantiques très différentes de /e/,voyelle antérieure, et de /o/, voyelle postérieure, nous ont amené à définir deux indices différents pour distinguer les réalisations diphtonguées des réalisations non-diphtonguées de ces deux voyelles.

Pour la voyelle /e/, la variation fréquentielle du deuxième formant s'est montré un critère opérationnel pour catégoriser les réalisations de notre corpus.

On peut donc énoncer la règle suivante: si l'amplitude de la variation de F2 au cours de la voyelle /e/ est supérieure à 400Hz, la réalisation de la voyelle est perçue comme étant diphtonguée.

Pour la voyelle /o/, l'amplitude de la variation du centre de gravité tel que nous l'avons calculée a permis d'identifier les réalisations perçues comme étant marquées. Les réalisations non-diphtonguées des locuteurs américains et les réalisations des locuteurs français ne sont pas distinguées.

On peut donc énoncer la règle suivante: si l'amplitude de la variation du centre de gravité au cours de la voyelle /o/ est supérieure à 100-120Hz, la réalisation de la voyelle est perçue comme étant diphtonguée. Les variations de la fréquence fondamentale n'ont pas pu être mises en relation directe avec la perception de la diphtongaison.

Nous souhaitons maintenant valider l'opérationnalité de ces critères en multipliant les observations, en faisant varier les voyelles, les contextes consonantiques et les locuteurs. Des tests d'évaluation plus élaborés (testant des stimuli dont les durées et les variations de la fréquence fondamentale seraient contrôlés) permettraient d'affiner la définition de seuils acoustiques correspondant à la perception de marques anglophones. Un système d'aide à la prononciation, sur lequel une telle modélisation serait implantée, serait susceptible de pouvoir évaluer les réalisations des voyelles françaises /e/ et /o/ prononcées par un apprenant américain à partir d'énoncés de référence. L'utilisateur d'un tel système pourrait donc recevoir un retour d'information pour apprécier le degré de diphtongaison de sa réalisation.

Remerciements

Cette étude est financée par une allocation de recherche MRT accordée au DEA de Phonétique de Paris. Remerciements à Samir CHENNOUKH et René CARRE pour avoir permis l'accès à leur Analyseur de signaux.

Références

CALLIOPE, (1989), *La parole et son traitement automatique*, Masson.

DELATTRE, Pierre, (1951), *Principes de Phonétique: A l'usage des étudiants anglo-américains*, Middleburry College, Vermont, Second Edition, .

DELATTRE, Pierre, (1965), *Comparing the Phonetic Features of English, French, German and Spanish*, Julius Groos Verlag, Heidelberg.

FLEGE, James Emil, (1984), The detection of French accent by American listeners, *JASA* 76(3), pp. 692-707.

FRY, D. B. , (1979), *The physics of speech*, Cambridge University Press.

GAY, Thomas, (1968), Effect of speaking rate on vowel formant movements, *JASA* 44(6), pp. 1570-1573.

GIMSON, A.C., (1962), *An inroduction to the pronunciation of English*, London, Edward Arnold (publishers) Ltd.

HEFFNER, R-M. S., (1964), *General Phonetics*, The University of Winconsin Press, Madison.

HERTZ, Susan R. , Streams, phones and transitions: toward a new phonological and phonetic model of formant timing, *Journal of Phonetics*, 1991, 19, p. 91-109.

LADEFOGED, Peter, (1975), *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc. , New York, pp. 63-82.

LADO, Robert, (1957), *Linguistics across cultures: applied linguistics for language teachers*, Ann Arbor, The University of Michigan Press.

LE BRAS, Jane, (1981), *Utilisation de la reconnaissance automatique de la parole pour l'apprentissage des langues*, Thèse, Rennes II.

LEHISTE, Ilse, PETERSON, Gordon E. , (1967), Transitions, Glides, and Diphtongs, in *Readings in acoustic phonetics*, edited by Ilse LEHISTE, The M.I.T. Press.

LEON, Pierre R., (1963), Perception of French vowels by American-english listeners, *Etudes de Linguistique appliquée*, Didier, 2, pp. 135-146.

MUNRO, Murray J., (1993), Productions of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings, *Language and Speech*, 36(1), pp. 39-66.

PETERSON, G.E., COXE, M.S. , (1953), The vowels /e/ and /o/ in American speech, *The quartely Journal of Speech* 39, p. 33-41.

TRANEL, Bernard, (1987), *The sounds of French: an introduction*, Cambridge University Press.

LES CONSONNES "FORTES" DU BERBÈRE : INDICES PERCEPTUELS ET CORRÉLATS PHONÉTIQUES

Naïma LOUALI ET Gilbert PUECH

UMR *"DYNAMIQUE DU LANGAGE"* - CNRS ET UNIVERSITÉ LUMIÈRE-LYON 2

RÉSUMÉ

Le système consonantique du berbère inclut une opposition *lenis / fortis*. Les consonnes de la série *fortis* sont considérées comme "longues", "géminées" ou "tendues". Cette contribution porte sur les corrélats phonétiques de cette opposition pour les obstruantes dans les dialectes tachlhit (Maroc) et touareg (Niger). Les paramètres retenus dans cette étude ont été : la durée, au plan acoustique et perceptuel, la pression intra-orale et le débit d'air buccal, l'effet sur le timbre des voyelles adjacentes. Il ressort des tests perceptuels effectués que la durée constitue l'indice majeur de discrimination mais que le jugement des locuteurs prend aussi appui sur des indices complémentaires.

LE SYSTÈME CONSONANTIQUE BERBÈRE

Il s'articule autour de trois corrélations : le voisement, l'emphase[1] et la "tension". En ce qui concerne la tension, toutes les consonnes berbères présentent l'opposition *lenis / fortis* ; toutefois l'homologue *fortis* de la consonne ḍ est l'occlusive pharyngalisée sourde ṭṭ et celui de ɣ est l'occlusive uvulaire sourde qq.

Une consonne *fortis* peut être interne au thème lexical ou résulter d'une jonction morphologique. Voici des exemples de paires minimales dans le parler de Tiznit au sud Maroc (dialecte tachlhit) pour des thèmes lexicaux non affixés (les consonnes *fortes* sont notées par une gémination) :

ks| ~ |kks|	[ks] ~ [kks]
"conduis au pâturage !" / "enlève !"	
|ili| ~ |illi|	[ili] ~ [illi]
"être" / "ma fille"	
|as| ~ |ass|	[as] ~ [ass]
"jour" / "attache !"	

1. L'emphase se définit articulatoirement comme la surimposition d'une articulation secondaire (il s'agit ici de la rétraction de la racine de la langue qui entraîne ainsi la constriction de la cavité pharyngale) à une articulation primaire. En berbère, le trait de pharyngalisation est phonologiquement associé aux consonnes apicales : /ḍ, ẓ, ṣ/ parfois /ṭ/ et/ou /ɾ/. Ce trait peut s'étendre à d'autres segments (consonantiques et vocaliques) à l'intérieur d'un même mot. Pour plus de précisions voir Louali et Puech (1989) et Louali (1990).

Ici, on trouve l'opposition à l'initiale, en médiane et en finale. Pour d'autres parlers, ainsi en touareg du Niger, l'opposition en initiale absolue n'est pas attestée.

Une consonne *fortis* peut résulter de la jonction morphologique entre une consonne radicale et une consonne affixale :

Formation du pluriel

[afus] → |i+fas+s+ən| [ifassən]
"main" / "mains"
[afud] → |i+fad+d+ən| [ifaddən]
"genou" / genoux"

Formation du diminutif

[afud] → |t+afud+t| [tafutt]
"genou" / "petit genou"

Formation de l'intensif

[səɣ] → |s+saɣ| [ssaɣ]
"achète!" / "achète souvent !"

Assimilation du morphème [ad]

|ad≠t+ddu| [attəddu]
"elle va venir"
|ad≠d#ddu+x| [addəddux]
"je vais venir vers ici"

ÉTUDES ANTÉRIEURES

Très tôt, les berbérisants se sont intéressés à la spécification de cette corrélation : s'agit-il de consonnes longues ou de consonnes tendues ? Basset (1952) adopte la terminologie de "consonnes longues". Mais, dès 1953, le terme de tension fait son apparition bien que ce soit sous forme interrogative : "Qui sait si la tension ne joue pas le rôle principal ?" (Galand, 1953, p. 230). En 1957, Mitchel décrit ce type de consonnes comme tendues en s'appuyant sur une étude palatographique du parler zwara (nord Libye). Galand (1960 et 1988) insiste sur la pertinence du caractère tendu de ces consonnes : en particulier, on ne peut pas rendre compte de l'alternance ḍ / ṭṭ et ɣ / qq en se référant au seul contraste de longueur.

A partir de données acoustiques du kabyle, Chaker (1975) conclut que la durée est le principal corrélat acoustique mais n'exclut pas le caractère tendu de ces consonnes. D'autres études telles que Saïb (1977) et Guerssel (1977), s'inspirant du modèle génératif, posent le problème en d'autres termes et s'intéressent davantage à la question de la représentation de ces consonnes comme *un* ou *deux* segments. Louali (1983) propose une approche syllabique de la tension en s'inscrivant dans le cadre théorique de la phonologie non linéaire, notamment celle préconisée par Clements et Kayser (1983). Dell et Elmedlaoui (1988) ont également étudié les problèmes posés par la représentation des consonnes et le traitement de la syllabe en berbère.
A ce stade de la réflexion sur le fonctionnement de ces consonnes et leur représentation, il est opportun de mettre en perspective l'analyse phonologique avec des données phonétiques, si l'on veut échapper à des schémas d'analyse circulaires où l'on s'appuie sur l'insertion du schwa[2] pour justifier la représentation des consonnes tendues et sur les mêmes représentations pour expliquer les règles d'insertion du schwa.
Pour cerner la nature de ces consonnes, nous les avons soumises à trois types d'analyses : perceptuelle, physiologique et acoustique.

ÉTUDE PERCEPTUELLE

Pour l'analyse perceptuelle, nous avons orienté les tests sur la durée des consonnes occlusives sourdes et sonores des parlers de Tiznit, de Tafraout et d'Inezgane (dialecte tachlhit). A partir d'un corpus plus large, nous avons retenu les séquences suivantes parce qu'elles étaient identifiées et réalisées de manière similaire par la plupart des locuteurs de cette zone dialectale :

[a g g a r]	"le fait de semer" / "semence"
[a g a r]	"criquet"
[i t t ə l]	"il s'est couvert"
[i t ə l]	"il a taché"

La préparation des stimuli consiste à prendre le signal correspondant à la réalisation de la consonne considérée et à repérer dans la partie stable du signal une période ; pour la consonne tendue, on diminue la durée par paliers de *x* ms jusqu'à atteindre la durée d'une non tendue ; on procède inversement par augmentation progressive de la durée lorsqu'on part de la consonne non tendue. Pour la paire [a g g a r] / [a g a r] , nous avons diminué et augmenté le signal par paliers de 10 ms. En ce qui concerne la paire [i t t ə l] / [i t ə l], elle a été traitée par paliers de 5 ms. Nous avons obtenu chaque fois neuf stimuli en plus du stimulus original. Les dix stimuli sont présentés au sujet dix fois chacun dans un ordre aléatoire. Le sujet avait à émettre un jugement sémantique et un choix forcé, c'est-à-dire que son attention est d'une part attirée uniquement sur le sens de ce qu'il entend et d'autre part limitée par un choix à deux possibilités. Les résultats sont analysés sous forme de graphique : l'axe horizontal représente la durée en ms, et l'axe vertical représente le nombre de jugements d'identification pour chacune des deux séries de présentation (mot avec consonne tendue abrégée par paliers, et mot avec consonne non tendue allongée).
En ce qui concerne la paire [a g g a r] / [a g a r] (figure 1), le croisement des deux courbes se fait autour de 140 ms. L'augmentation du signal de 60 ms dans un cas et sa diminution de 50 ms dans l'autre ne perturbe en rien les stratégies perceptuelles du sujet qui continue à distinguer les deux mots de la paire à 100% jusqu'à 140 ms, un seuil au delà duquel son jugement bascule et provoque une chute du taux de reconnaissance dans un premier temps autour de 60% puis 35% et enfin 0%. Ce graphique montre évidemment que la durée est un indice qui guide le sujet mais surtout qu'il n'est pas le seul : il est clair que ce sujet s'appuie dans ces jugements sur d'autres indices que la durée de la consonne dans la mesure où il arrive à discriminer de manière spectaculaire les deux mots de la paire malgré le taux important de la réduction ou de l'augmentation de la durée de la consonne.
La figure 2 regroupe les résultats obtenus pour trois sujets. Le croisement des courbes représentant les séquences [a g g a r] / [a g a r] modifiées par paliers se produit autour de 140 ms. Pour ce seuil critique de durée, on obtient 50% d'identification conforme au stimulus dont on est parti. Le graphique confirme que pour ces trois sujets la durée n'est pas le seul critère discriminatoire. Il est secondé par d'autres indices qui leur permettent de maintenir distincts les deux mots de la paire à 30 ms d'augmentation ou de réduction de la consonne. La baisse de taux de reconnaissance se fait de manière progressive et on ne note une chute qu'au delà et en deçà de 140 ms.
Pour les occlusives sourdes, la figure 3 représentant les séquences [i t t ə l] / [i t ə l] montre que pour trois sujets le croisement des courbes se fait autour de 190 ms et que la baisse de taux de reconnaissance est plus progressive. Ces données attestent que les berbérophones, pour distinguer [g g] / [g] ou [t t] / [t] , utilisent des stratégies perceptuelles qui font recours à la durée comme indice principal mais qu'ils s'appuient subsidiairement sur des indices complémentaires.
Quels sont ces indices, et comment se manifestent-ils au niveau physiologique et acoustique ?

2. Nous rappelons que pour les parlers berbères du nord, le schwa, voyelle centrale [ə], apparaît pour rompre une suite de consonnes. Dans cette investigation il faut prendre en considération à la fois le fonctionnement de la syllabe, la variation dialectale mais aussi la variation inter-locuteurs.

ÉTUDE PHYSIOLOGIQUE

L'étude physiologique porte sur la comparaison des productions des occlusives du parler de Tiznit (dialecte tachlhit) : /b/ ~ /bb/, /t/ ~ /tt/, /d/ ~ /dd/, /k/ ~ /kk/ et /g/ ~ /gg/. Le corpus[3] a été enregistré à l'aide d'un aerophonomètre[4] réalisé à l'Institut de Phonétique d'Aix-en-Provence[5]. L'objectif ici est de comparer les courbes qui correspondent à la pression buccale (P.B) et la pression intra-orale (P.I.O) lors de la production d'une consonne tendue et de sa contre partie non tendue.

La figure 4 représente la production des séquences [ittəl] / [itəl]. La comparaison des courbes de la P.I.O met en évidence d'une part la durée comme paramètre distinctif et d'autre part une différence assez nette des profils des deux courbes. Quant aux courbes qui reflètent la P.B, la différence de la consonne tendue est traduite essentiellement par une montée rapide et importante au moment de l'explosion. Les mêmes configurations des courbes, profil différent (pour la P.I.O) et montée rapide et importante au moment du relâchement (pour la P.B), sont relevées lors de la comparaison des occlusives tendues sonores aux occlusives non tendues sonores, comme le montre la figure 5.

Les indices que constituent les profils différenciés des courbes de P.I.O et de P.B confirment la présence de co-facteurs dans la détection de la tension des consonnes du berbère. Nous allons voir maintenant comment ces corrélats physiologiques se traduisent au niveau acoustique.

ÉTUDE ACOUSTIQUE

Les données acoustiques proviennent des parlers de Tiznit (dialecte tachlhit) et de celui de Gofat (dialecte tayirt-touareg). La comparaison des sonagrammes 1 et 2 (figure 6) montre d'une part que la durée de la consonne tendue est plus importante que sa contre-partie non tendue et d'autre part que la qualité de l'énergie de l'explosion est plus intense dans la zone de fréquence entre 1000 et 2000 Hz. Quant au sonagramme 4 (figure 7), il met bien en évidence la durée mais surtout le dévoisement sur la fin de la consonne tendue voisée.

Il ressort d'une comparaison, dont les résultats détaillés ne pouvaient être inclus dans cette présentation, que dans le parler de Tiznit la durée moyenne des consonnes tendues est au moins de 50% plus importante que celle des consonnes non tendues. Quant à la durée moyenne des consonnes tendues du parler de Gofat, elle est nettement supérieure à celle des consonnes non tendues, mais son pourcentage varie d'une consonne à l'autre.

La durée de l'explosion des consonnes sourdes est significative dans le cas du parler de Gofat. Pour le parler de Tiznit en revanche, cette distinction fonctionne seulement pour les vélaires.

Les analyses effectuées pour mesurer l'effet acoustique de ces consonnes sur leur entourage vocalique, nous permettent de conclure que l'influence des valeurs formantiques n'est pas significative pour ces deux parlers. Cependant les données montrent une tendance à l'abrégement de la voyelle qui précède ou qui suit la consonne tendue. Nous avons noté que la réduction de la durée de la voyelle qui précède la consonne tendue est plus importante que celle de la voyelle qui suit.

Ainsi, l'analyse acoustique met en évidence trois corrélats associés à la réalisation des consonnes *fortes* : la durée, la qualité de l'énergie de l'explosion des consonnes sourdes et le dévoisement partiel des consonnes sonores. De même que la longueur des voyelles adjacentes est réduite au contact des consonnes tendues.

CONCLUSION

Le test perceptuel montre que pour ce type de consonnes la durée est un indice principal mais que les locuteurs s'appuient sur d'autres indices secondaires dans leurs jugements. Les études physiologique et acoustique des données confirment l'existence de corrélats susceptibles de constituer des indices complémentaires sur lesquels reposent les stratégies perceptuelles de catégorisation des locuteurs.

3. Voir Annexe.

4. Cet appareil qui permet d'enregistrer simultanément au cours de la production le signal, les vibrations du larynx, les débits d'air au niveau des narines, les pressions buccale et intra-orale.

5. Nous saisissons cette occasion pour remercier *Bernard Teston*, Ingénieur de Recherches au CNRS, pour avoir mis à notre disposition ce matériel et nous avoir assistés dans son utilisation.

RÉFÉRENCES

BASSET A. (1946), "Le système phonologique du berbère" *G. L. E. C. S.* (Groupe Linguistique des Etudes Chamito-Sémitiques), IV, pp. 33-36.

BASSET A. (1952), *La langue berbère*, published for the International African Institute by Dawson of Pall Mall, London.

BOUKOUS A. (1987), *Phonotactique et domaines prosodiques en berbère (parler tachelhit d'Agadir, Maroc)*, Thèse d'État, Université Paris VIII.

CHAKER S. (1975), "Les paramètres acoustiques de la tension consonantique en berbère" (dialecte kabyle), in *Travaux de l'Institut de Phonétique d'Aix*, pp. 151-168.

CLEMENTS G. N. & KAYSER S. J. (1983), *CV Phonology : A Generative Theory of the Syllabe*, MIT Press, Cambridge, London.

DELL F. & ELMEDLAOUI F. (1988), "Syllabic consonants in Berber: some new evidence", *Journal of African Languages and Linguistics*, 10,1-17.

GALAND L. (1953), "La phonétique en dialectologie berbère, *Orbis*, 2/1, pp. 225-233.

GALAND L. (1960), "Berbère", La Langue, *Encyclopédie de l'Islam, Phonétique et phonologie*, published by Brill, tome I, pp. 1216-1217.

GALAND L. (1988), "Le berbère", *Les langues dans le monde ancien et moderne, Les langues chamito-sémitiques*, 3ème partie, Paris C. N. R. S. , pp. 207-242.

GUERSSEL M. (1977), *Issues in Berber phonology*, Ph. D. Thesis, University of Washington.

KOHLER K.J. (1982), "FØ in the production of lenis and fortis plosives", *Phonetica*, 39, p.199-218.

LOUALI N. (1983), *Les corrélations consonnantiques en berbère : parler d'El Aderj* (Moyen Atlas, Maroc), Mémoire de D. E. A. de linguistique, Université Lumière-Lyon 2.

LOUALI N. & PUECH G. (1989), "La pharyngalisation des consonnes labiales en berbère", *J. Acoustique* , n° 2, pp. 147-154.

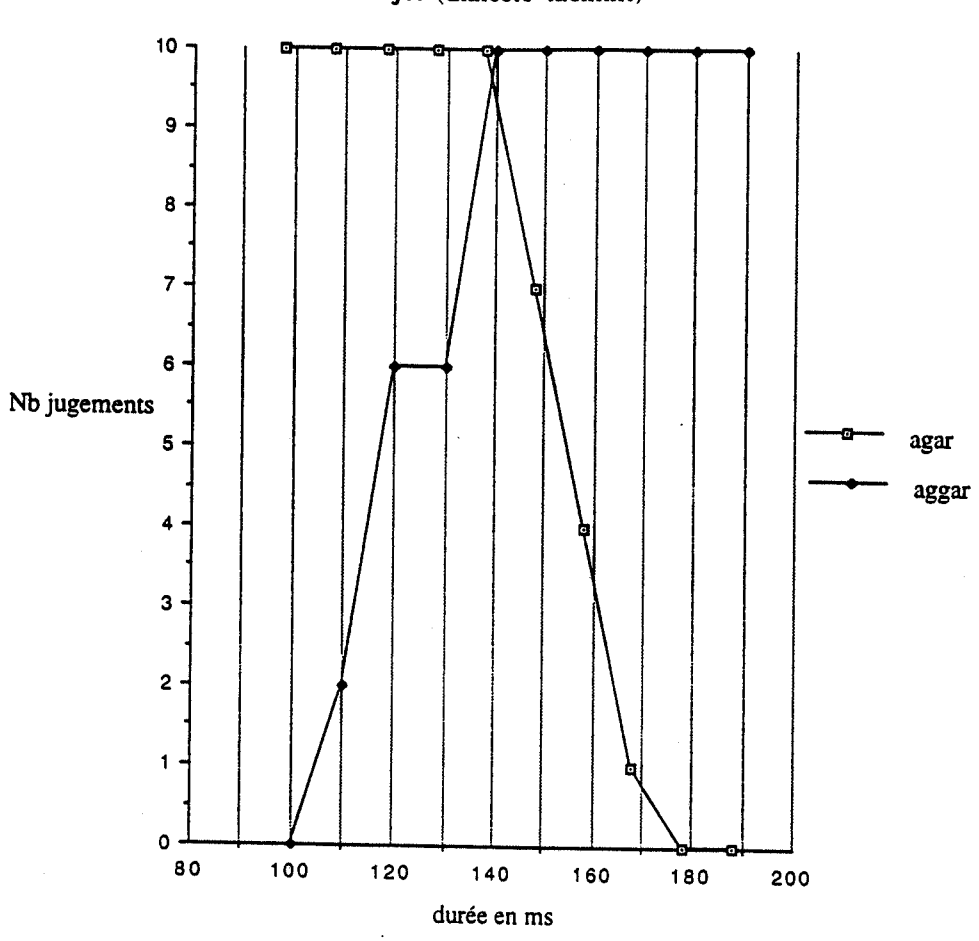
LOUALI N. (1990), *L'emphase en berbère. Etude phonétique, phonologique et comparative*, Thèse de doctorat, Université Lumière-Lyon

MALÉCOT A. (1970), "The lenis-fortis opposition: its physiological parameter", *Journal of the Acoustical Society of America*, 47, 6 (part 2).

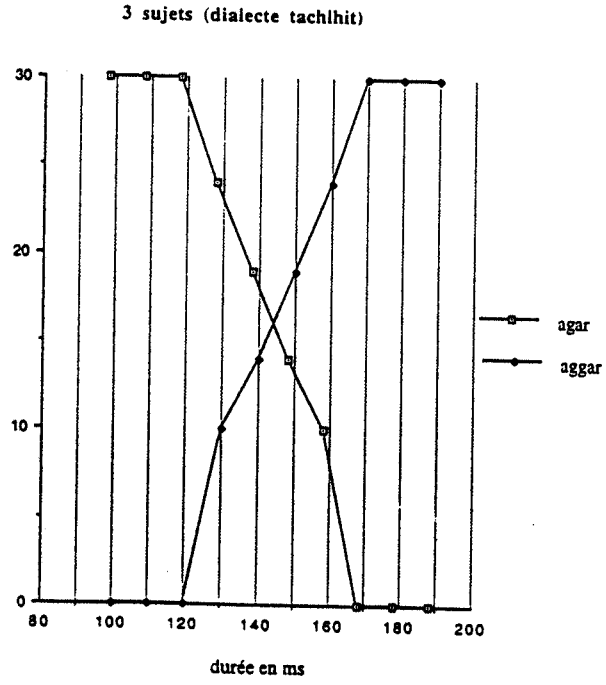
MITCHELL T.F. (1957), "Long Consonants in Phonology and Phonetics", *Studies in Linguistic Analysis*, Philological Society, Oxford University.

SAÏB J. (1977), "The Treatment of Geminates : Evidence from Berber", *Studies in African Linguistics*, volume 8, n° 3, pp. 299-316.

Figure 1. Résultats d'un test perceptuel sur la durée d'une consonne *lenis* et d'une consonne *fortis* du berbère.



Figures 2 et 3. Résultats d'un test perceptuel sur la durée d'une consonne *lenis* et d'une consonne *fortis* du berbère.



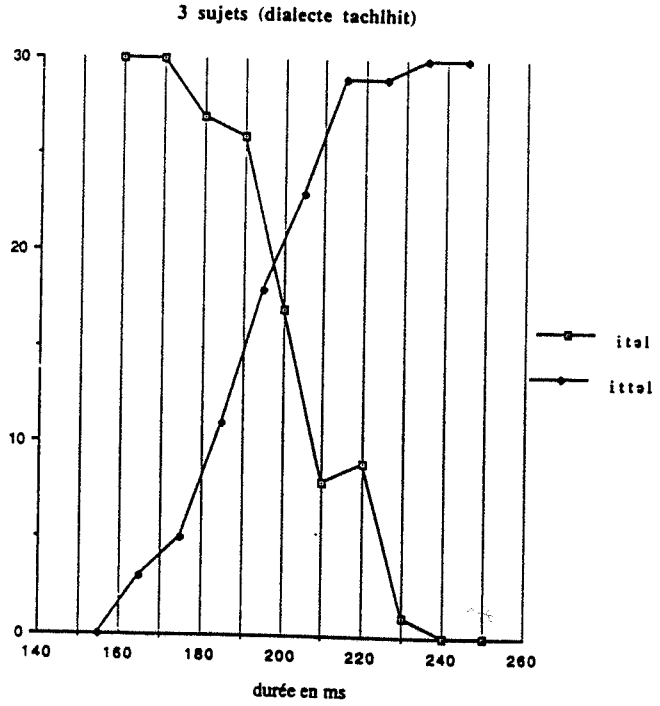


Figure 4. Réalisation des séquences [ittəl]/[itəl]

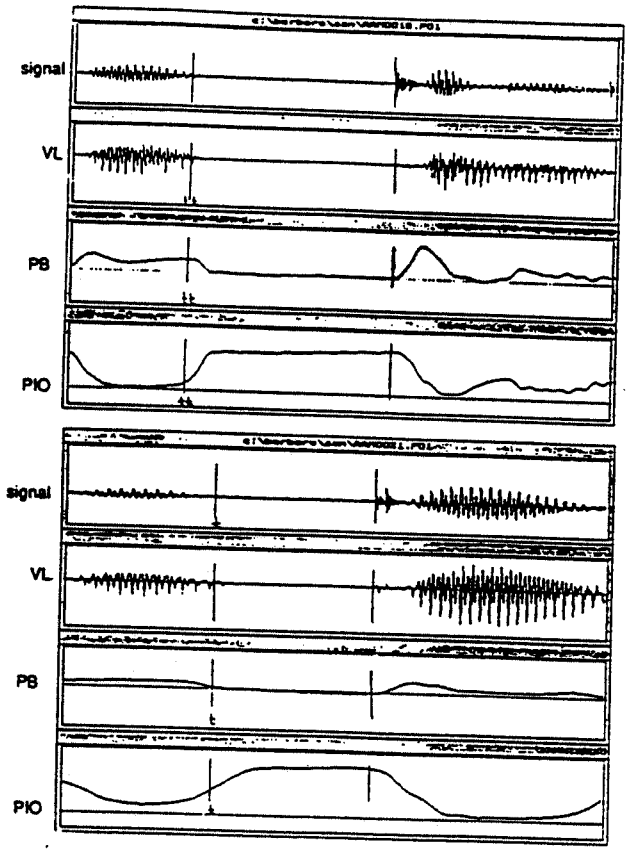
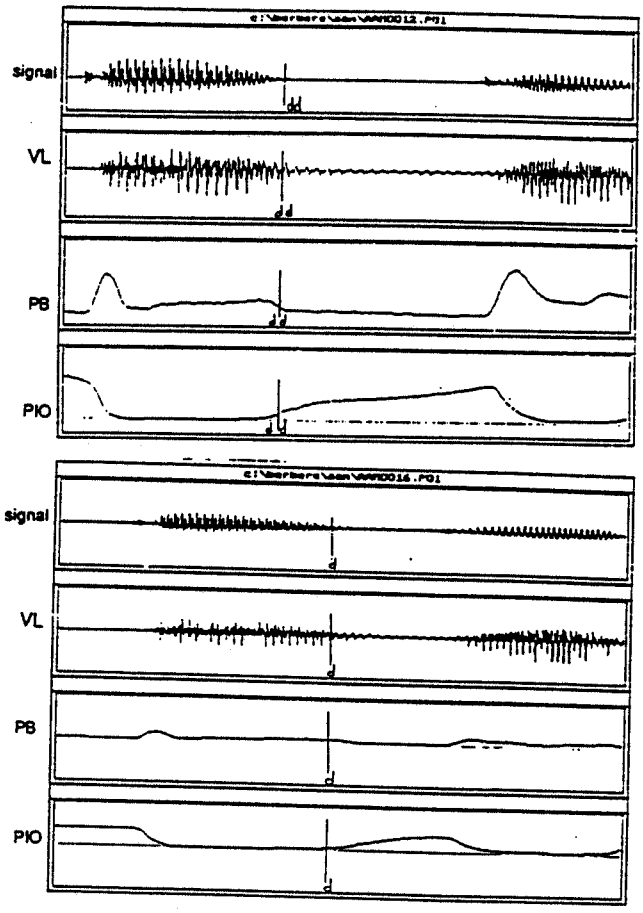


Figure 5. Réalisation des séquences [tajdditt]/[tajdit]



Annexe : Corpus

parler de tiznit-tachlhit (Sud-Maroc)

addal	sorte de cape
adan	intestins
tajddit	petite outre
tajdit	chienne
ittəl	il s'est couvert
itəl	il a taché (ses mains etc.)
tittəlt	le fait de se couvrir
titəlt	le fait de se tacher
tukkimt	poing
tikint	marmite
aggar	le fait de semer / semence
agar	criquet
iggər	il a touché
igər	champ

parler de gofat-tayert (touareg-Niger)

xabba	prénom mas.
xaba	prénom fem.
addal	jeu
adal	guépard
əddər	vivre
ədər	se calmer
attəl	envelopper
atəl	emporter (affaires/troupeau)
aggar	fruit du tanin
agar	marea grassifolia
əggəd	sauter
əgət	une partie d'un tronc d'arbre mort
aggad	le fait d'attendre
agad	bruit très fort
akkal	souffrance
akal	pays

Figure 6. Sonagrammes des séquences [i t ə l] et [i t t ə l]

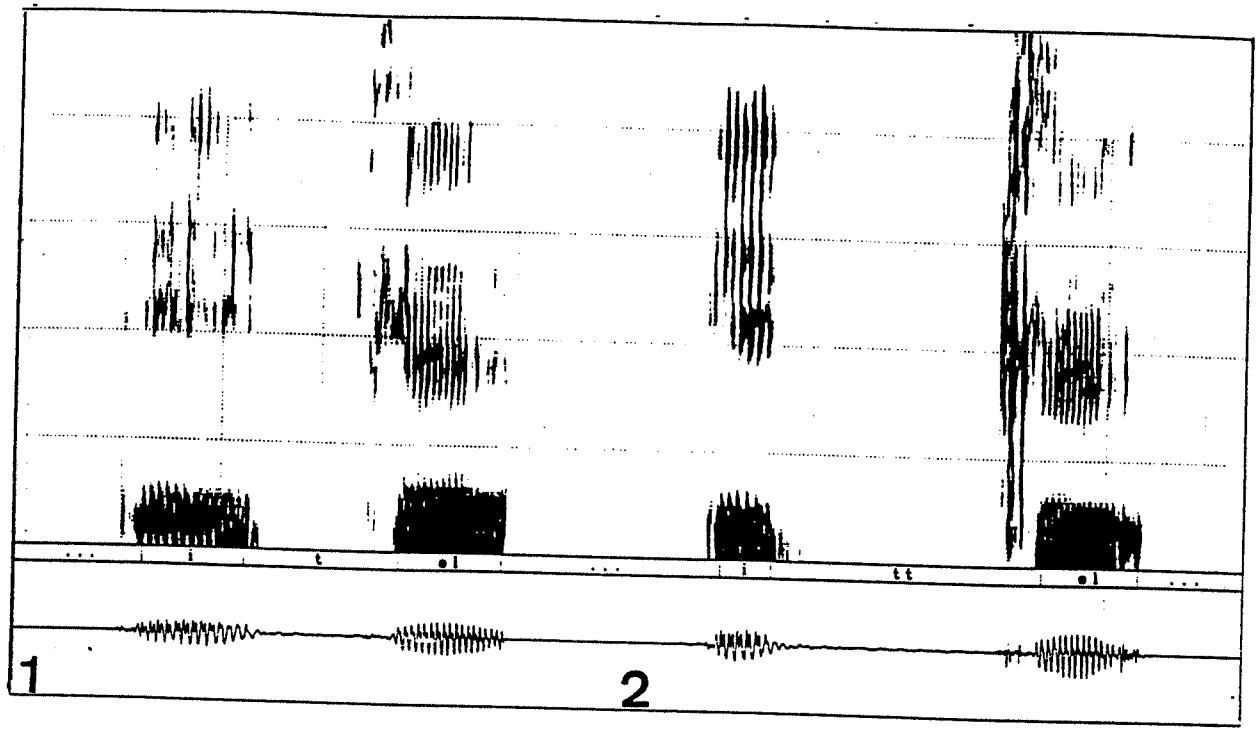
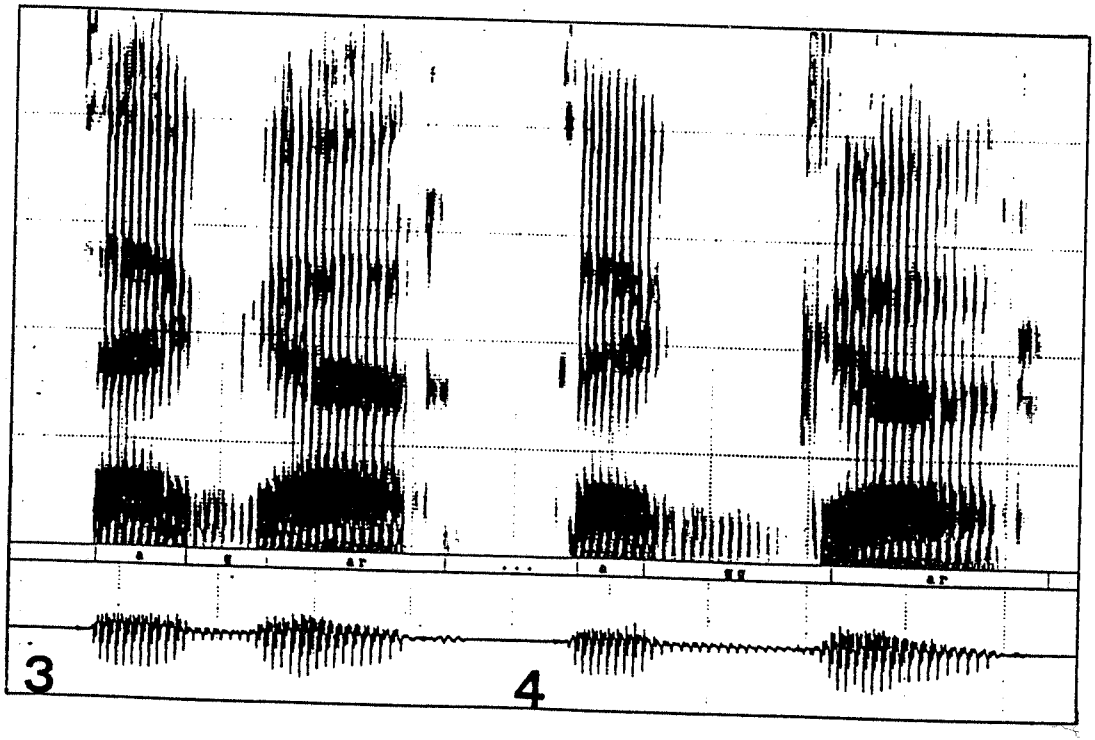


Figure 7. Sonagrammes des séquences [a g a r] et [a g g a r]



XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O6

RECONNAISSANCE

MODELES DE SEQUENCES A HORIZON VARIABLE : MULTIGRAMS.

Frédéric BIMBOT (2) (1), Roberto PIERACCINI (1), Esther LEVIN (1) et Bishnu ATAL (1).

(1) AT&T Bell Laboratories - Speech Research Dept, 600 Mountain avenue, MURRAY HILL - NJ 07974, USA.
(2) ENST - dept SIGNAL, CNRS - URA 820, 46 rue Barrault, 75634 PARIS cedex 13, FRANCE, Union Européenne.

ABSTRACT

Spoken and written language results from a complex encoding of a message into a stream of acoustic events or symbolic units. This process is highly redundant. One can therefore expect to retrieve, in an observation string, typical variable-length sequences of elementary units, which express articulatory, lexical and syntactic constraints. This article presents the multigram model, which aims at parsing a string of symbols into variable-length independent sequences. We illustrate the result of the model on the decomposition of character sequences into sub-word units, and we evaluate its performance for a language modeling task.

Mots-clefs : séquences de longueur variable, modèle probabiliste de langage.

I - INTRODUCTION

Le langage parlé peut être considéré comme le résultat d'un processus complexe d'encodage d'une idée en une suite de phonèmes, dont la réalisation orale résulte en une onde acoustique. Ce processus d'encodage est néanmoins soumis à différentes contraintes : des contraintes de production garantissant qu'une suite de phonèmes soit prononçable, des contraintes lexicales liées à la limitation du vocabulaire et des contraintes syntaxiques régissant l'ordre des mots. Dès lors, toutes les combinaisons possibles de phonèmes ne sont pas observées dans la chaîne parlée, loin s'en faut.

Il existe donc une certaine redondance du processus d'encodage phonétique qui a pour conséquence une indéniable robustesse des mécanismes de la communication parlée. D'un point de vue pratique, cette redondance se traduit par l'existence de séquences typiques d'événements acoustiques, traduisant la régularité du lexique et, dans une moindre mesure, les contraintes syntaxiques. Ces séquences sont naturellement de longueur variable.

L'objet de cet article est de présenter un modèle de séquences à horizon variable permettant de rendre compte des régularités dans la structure du langage parlé et écrit. Nous exposons les principes de ce modèle de "multigrams", et nous illustrons son fonctionnement sur 2 exemples : d'une part, la recherche de séquences typiques dans un texte écrit dont les espaces entre les mots ont été supprimés et d'autre part la modélisation probabiliste de l'enchaînement entre les mots dans un corpus de texte. Le premier exemple conduit à la définition d'unités segmentales intermédiaires entre le caractère et le mot. Le second démontre l'intérêt de l'approche en modélisation du langage.

II - GENERALITES

Le modèle de multigrams présenté dans cet article peut être compris par opposition au modèle classique de n-grams, présenté de façon très détaillée dans [1].

Modèle de n-grams :

Le modèle classique de n-grams, dans sa version la plus simple, estime la vraisemblance d'une chaîne de N symboles :

$$W = w_1 w_2 \ldots w_N$$

par la formule :

$$L(W) = L(w_1 w_2 \ldots w_N) = \prod_{i=1}^{N} P(w_i / w_{i-n+1} \ldots w_{i-1})$$

où $P(w_i / w_{i-n+1} \ldots w_{i-1})$ est la probabilité conditionnelle d'apparition du symbole w_i connaissant les n-1 symboles précédents.

Par exemple, pour une chaîne de 4 symboles et un modèle de tri-grams :

$$\begin{aligned} L(W) &= L(w_1\, w_2\, w_3\, w_4) \\ &= P(w_1 / \#\,\#) \,.\, P(w_2 / \#\, w_1) \,. \\ &\qquad P(w_3 / w_1\, w_2) \,.\, P(w_4 / w_2\, w_3) \end{aligned}$$

où # indique le début de chaîne.

L'hypothèse sous-jacente à une telle approche est que la probabilité d'un symbole donné dépend exactement des n-1 symboles précédents, ce qui est clairement une hypothèse trop simpliste pour le langage naturel.

III - LE MODELE DE MULTIGRAMS

Par opposition au modèle de n-grams, le modèle de multigrams que nous proposons peut être compris comme un modèle probabiliste pour lequel un symbole dépend d'un contexte à horizon variable. Une chaîne de symboles est décomposée en séquences de longueur variable et indépendantes entre elles, cette segmentation s'effectuant au sens du maximum de vraisemblance.

Soit S_k une sous-chaîne de la chaîne $W = w_1\, w_2 \ldots w_N$, commençant à l'indice i_k et finissant à l'indice $i_{(k+1)}-1$, c'est-à-dire :

$$S_k = [w_{i_k}\, w_{i_k+1} \ldots w_{i_{(k+1)}-1}]$$

La segmentation de W en q séquences peut être décrite par un ensemble B d'indices :

$$B = \{i_1, i_2, \ldots, i_q, i_{(q+1)}\}$$

marquant le premier symbole de chaque séquence, avec les contraintes suivantes :

$$i_1 = 1 \text{ et } i_{(q+1)} = N+1$$
$$i_1 < i_2 < \ldots < i_{(q+1)}$$

Si de plus on restreint à n la longueur maximale de chaque séquence, on aboutit à la contrainte additionnelle :

$$\text{pour } 1 \le k \le q, \quad 1 \le i_{(k+1)} - i_k \le n$$

Si l'on note finalement {B} l'ensemble de tous les ensembles B satisfaisant les contraintes ci-dessus, le modèle de n-multigrams estime la vraisemblance d'une séquence W par la formule :

$$L(W) = \max_{\{B\}} \prod_{k=1}^{q} P(S_k)$$

Ainsi, pour une chaîne de 4 symboles et un modèle de 3-multigrams :

$$\begin{aligned} L(W) &= L(w_1\, w_2\, w_3\, w_4) \\ &= \max \{\ P([w_1\, w_2\, w_3])\, P([w_4]), \\ &\qquad P([w_1])\, P([w_2\, w_3\, w_4]), \\ &\qquad P([w_1\, w_2])\, P([w_3\, w_4]), \\ &\qquad P([w_1\, w_2])\, P([w_3])\, ([w_4]), \\ &\qquad P([w_1])\, P([w_2\, w_3])\, P([w_4]), \\ &\qquad P([w_1])\, P([w_2])\, P([w_3\, w_4]), \\ &\qquad P([w_1])\, P([w_2])\, P([w_3])\, P([w_4])\ \} \end{aligned}$$

La décomposition en séquences qui maximise la fonction L fournit une segmentation de la chaîne W.

IV - SEGMENTATION

Si l'on suppose connues les probabilités de toutes les séquences possibles de symboles w_k, une chaîne W peut être décomposée par un algorithme de Viterbi.

Ainsi, pour toute sous-chaîne $w_1 \ldots w_{k+1}$ de W jusqu'à l'indice k+1, la vraisemblance partielle $L(w_1 \ldots w_{k+1})$ pour un modèle de multigrams d'ordre n peut se calculer comme le maximum des n termes suivants :

$$\begin{aligned} &P([w_{k+1}])\ L(w_1 \ldots w_{k-n+1}\, w_{k-n+2} \ldots w_{k-2}\, w_{k-1}\, w_k), \\ &P([w_k\, w_{k+1}])\ L(w_1 \ldots w_{k-n+1}\, w_{k-n+2} \ldots w_{k-2}\, w_{k-1}), \\ &P([w_{k-1}\, w_k\, w_{k+1}])\ L(w_1 \ldots w_{k-n+1}\, w_{k-n+2} \ldots w_{k-2}), \\ &\qquad \ldots, \\ &P([w_{k-n+2} \ldots w_{k-2}\, w_{k-1}\, w_k\, w_{k+1}])\ L(w_1 \ldots w_{k-n+1}). \end{aligned}$$

c'est-à-dire :

$$L(w_1 \ldots w_{k+1}) = \max_{1 \le i \le n} \{P([w_{k-i+2} \ldots w_{k+1}])\ L(w_1 \ldots w_{k-i+1})\ \}$$

Cette formule est utilisée de façon récursive pour la segmentation de la chaîne W. La quantité L (W) obtenue pour la segmentation optimale fournit naturellement la vraisemblance de W au sens du modèle.

V - ESTIMATION DES PROBABILITES

Etant donné un corpus d'apprentissage constitué d'une (ou de plusieurs) chaîne(s) W, de longueur totale N, les probabilités des séquences pour le modèle de n-multigrams peuvent être estimées en 2 temps : une estimation initiale et une ré-estimation itérative. L'estimation initiale utilise toutes les séquences de symboles observées dans le corpus, alors que la ré-estimation des probabilités s'effectue à partir des séquences obtenues à l'itération précédente.

Estimation initiale :

Soit c_{init} (T) le nombre d'occurrences d'une sous-chaîne de symboles T dans le corpus d'apprentissage, et soit :

$$C_{init} \approx n N$$

le nombre total de sous-chaînes, toutes longueurs confondues (jusqu'à n).

L'estimation de la probabilité de la sous-chaîne T est obtenue par :

$$P_{init}(T) = \frac{c_{init}(T)}{C_{init}}$$

Ré-estimation itérative :

Les estimations initiales sont utilisées pour segmenter le corpus d'apprentissage selon l'algorithme du paragraphe IV. L'ensemble d'apprentissage est donc décomposé en un certain nombre C de séquences de longueur inférieure ou égale à n. Si l'on note c (S) le nombre de séquences S dans le corpus segmenté, on peut ré-estimer la probabilité de la séquence S par :

$$P(S) = \frac{c(S)}{C}$$

Ces nouvelles probabilités permettent de re-segmenter le corpus d'apprentissage, et le processus est itéré jusqu'à convergence, ou pour un nombre d'itérations fixé au préalable.

La différence existant entre l'estimation initiale et la ré-estimation itérative réside en la façon d'estimer la probabilité d'une séquence. Parce que les données initiales ne sont pas segmentées, la phase initiale utilise toutes les sous-chaînes de symboles dans l'ensemble d'apprentissage (jusqu'à la longueur n) comme hypothèse de départ. Au contraire, les phases de ré-estimation ne prennent en compte que les séquences trouvées lors de l'itération précédente, en général en nombre bien moindre que pour l'étape initiale.

A l'issue du traitement itératif présenté ci-dessus, on obtient un dictionnaire de séquences de longueur variable permettant de décrire le corpus d'apprentissage.

Pénalisation des séquences rares :

Afin de réduire le nombre de séquences dans le dictionnaire et d'augmenter la capacité de généralisation du modèle sur un corpus test, nous avons observé qu'il pouvait être utile de modifier l'estimation des probabilités des séquences.

Une façon de faire consiste à remplacer l'estimation usuelle de la probabilité basée sur la fréquence d'occurrence, par la borne inférieure de l'intervalle de confiance de cette estimation. Ainsi, au lieu de calculer la probabilité d'une séquence S par la formule :

$$P(S) = \frac{c(S)}{C}$$

on peut calculer :

$$P'(S) = \frac{c(S)}{C}\left(1 - a\sqrt{\frac{C - c(S)}{C \cdot c(S)}}\right)$$

Pour a = 1.96, P' s'interprète comme la borne inférieure de l'intervalle de confiance à 95 %. Naturellement, quand P' < 0, on pose P' = 0. L'ensemble de ce calcul conduit à des probabilités dont la somme est inférieure à 1, mais une renormalisation des valeurs est possible. Dans la suite du texte, nous désignons par "coefficient de pénalisation" le paramètre a.

P' peut être compris comme une estimation de P évaluée "dans le pire des cas". Le rapport P' / P est d'autant plus proche de 1 que la séquence considérée est fréquente et donc que l'estimation de P est fiable. Les séquences rares sont plus pénalisées que les séquences fréquentes.

VI - ILLUSTRATION

L'exemple qui suit illustre les résultats obtenus par le modèle de multigrams pour retrouver des séquences typiques de lettres dans un texte dont les espaces entre les mots ont été intentionnellement supprimés. En effet, l'existence d'un lexique pour constituer le texte d'origine entraine la présence de successions non-aléatoires de symboles au sein du texte sans espaces.

Pour cet exemple, nous avons utilisé les Psaumes du Vieux Testament en anglais (King James version). Ce corpus contient approximativement 200 000 caractères. L'alphabet de symboles consiste en 26 lettres minuscules (les majuscules ayant été remplacées par des minuscules, les symboles de ponctuation ayant été supprimés).

Le modèle utilisé est un modèle de 5-multigrams, et le coefficient de pénalisation a a une valeur de 1.96. Moins de 10 itérations ont été nécessaires pour obtenir la convergence de l'algorithme. Le dictionnaire de séquences final contient environ 1 100 unités.

La figure 1 présente le résultat obtenu pour les 11 premiers versets, sur l'ensemble d'apprentissage lui-même.

On observe clairement que des mots ou morphèmes typiques de l'anglais sont extraits par l'algorithme, plus particulièrement ceux qui sont fréquents. Il arrive que des mots courts soient groupés ensemble, notamment lorsqu'ils sont souvent associés syntaxiquement (inthe, ofthe, inhis, ...). A l'inverse, les mots rares sont décomposés en unités plus simples (se at, l ea f, ch a f f, ve x, ...). A l'occasion, une segmentation inappropriée mais parfois pertinente (voire poétique !) se produit (river sof, thew in d, so red is pleas ure pour "sore displeasure").

Il est rare de disposer d'un texte sans espaces entre les mots, que l'on souhaiterait segmenter a posteriori ! L'exemple présenté ici n'est qu'une illustration du type de résultats obtenus. En revanche, le même principe peut être utilisé pour organiser automatiquement un corpus de transcriptions phonétiques, ou d'événements acoustiques quantifiés, par exemple des cibles de décomposition temporelle [2] [3]. En traitement automatique de la parole, cela peut déboucher sur la création non-supervisée d'unités de synthèse ou de reconnaissance optimales, au sens du maximum de vraisemblance.

VII - EVALUATION

Pour estimer les potentialités du modèle de multigrams, nous l'avons évalué dans le cadre d'un modèle de langage. Il ne s'agit plus, dans ce paragraphe, de retrouver des unités de type sous-mot à partir de successions de caractères, mais des unités de type "groupe de mots" à partir de successions de mots. En effet, de même que les séquences de lettres (ou de phonèmes) sont conditionnées en premier lieu par le lexique, les séquences de mots sont contraintes principalement par la syntaxe (et la sémantique). Il existe donc des régularités dans les successions de mots, que les modèles classiques de n-grams exploitent, mais en utilisant un contexte de longueur fixe.

Nous comparons les 2 modèles en utilisant la perplexité comme mesure de leur efficacité respective. Alors que les symboles élémentaires étaient des caractères dans l'exemple précédent, il s'agit de mots dans ce paragraphe.

Perplexité :

On a coutume de mesurer l'efficacité d'un modèle probabiliste de langage par la perplexité du modèle sur un ensemble de test (non-utilisé pour l'estimation des probabilités). Cette quantité se définit par $X = 2^H$, où l'entropie H est estimée sur l'ensemble de test W_{test} par :

$$H = -\frac{1}{N} \log_2 (L(W_{test}))$$

En remarquant que :

$$H = -\frac{1}{N} \log_2 \left(\left[\frac{1}{X}\right]^N\right)$$

on peut interpréter la perplexité comme étant la taille d'un vocabulaire de mots équiprobables et indépendants, qui conduirait à une difficulté équivalente de prédiction par rapport au langage étudié lorsque ce dernier est muni de son modèle probabiliste. En d'autres termes, le modèle de langage réduit l'incertitude sur un mot donné connaissant son contexte, et plus la perplexité du modèle est faible, mieux il rend compte des particularités du langage.

Séquences non-observées :

Il peut se produire que, dans un corpus de test, une succession de symboles ne soit pas décomposable en séquences du dictionnaire. Il convient donc d'attribuer une probabilité faible mais non nulle, à certaines des séquences non-observées dans le corpus d'apprentissage, afin que toute chaîne de test soit segmentable.

Dans le cas présent, nous attribuons la valeur :

$$P_o(S) = \frac{1}{2N}$$

pour toutes les séquences de 1 symbole qui ont une probabilité nulle dans le dictionnaire. Les autres séquences non-observées conservent une probabilité nulle. Ainsi, toute chaîne de test pourra au pire être segmentée en séquences de 1 symbole aux endroits où aucune autre solution n'est possible.

Comparaison avec les n-grams :

Pour notre évaluation, nous comparons les performances du modèle de 5-multigrams (et plusieurs coefficients de pénalisation) avec les modèles classiques de n-grams, en termes de perplexité.

Pour les n-grams, une probabilité :

$$P_o(S) = \frac{1}{2N}$$

est affectée à tout n-gram de l'ensemble de test qui n'aurait pas été observé dans l'ensemble d'apprentissage.

En outre, les probabilités des n-1 n-grams en début de phrase sont approximées par celles des n-grams d'ordre inférieur. Par exemple, dans le cas de tri-grams :

$$P(w_1 / \# \#) = P(w_1)$$
$$P(w_2 / \# w_1) = P(w_2 / w_1)$$

Résultats :

Le corpus d'apprentissage est le texte retranscrit d'énoncés spontanés de la base de données ATIS (Air Travel Information System). Ce corpus comprend environ 10 000 phrases (soit à peu près 100 000 mots) prononcées par des sujets face à un dispositif simulant un système interrogeable en langage naturel ("Wizard-of-Oz"), délivrant des informations sur les vols intérieurs américains. Les noms de villes, de mois, de jours, de compagnies aériennes et les nombres sont remplacés par un mot unique pour chaque catégorie. Le vocabulaire ainsi obtenu contient environ 900 mots.

Le corpus de test est constitué de 1 000 autres phrases (soit environ 10 000 mots).

Dans la table 1, on compare les performances de différents modèles de langage, en termes de nombre d'unités et de perplexité sur l'ensemble de test. On donne également la valeur de la perplexité sur l'ensemble d'apprentissage, la différence entre les deux valeurs indiquant la capacité de généralisation du modèle. La première moitié de la table correspond aux scores obtenus avec des modèles de n-grams ($1 \leq n \leq 5$); la seconde moitié, à ceux relatifs à un modèle de multigrams de longueur maximale égale à 5 et pour plusieurs coefficients de pénalisation.

modèle	nb unités	test	appr.
---------- n-grams ----------			
1-grams	931	91.5	80.5
2-grams	7253	21.0	11.8
3-grams	16589	33.1	6.5
4-grams	24121	69.5	5.2
5-grams	27979	123.4	4.7
---------- 5-multigrams ----------			
a = 0.00	8928	27.6	6.9
a = 1.00	4328	18.4	9.3
a = 1.96	2256	17.7	13.0
a = 3.00	1569	20.5	15.6
a = 5.00	1218	26.0	19.3

Table 1 : résultats sur la base de données ATIS
Perplexités comparées des n-grams et des multigrams

Ces résultats montrent qu'avec 2 à 3 fois moins d'unités que pour les bigrams, les 5-multigrams (avec un facteur de pénalisation compris entre 1.00 et 2.00) permettent d'obtenir une perplexité plus faible, sur le corpus testé.

La figure 2 donne un exemple de la segmentation des 10 premières phrases du corpus de test. On observe que les frontières des groupes de mots obtenus ont tendance à coïncider avec des frontières de groupes syntaxiques ou avec des frontières de concepts.

VIII - CONCLUSIONS

Le modèle de multigrams part d'hypothèses différentes de celles du modèle classique de n-grams : une chaîne de symboles est décomposée en séquences de longueur variable, indépendantes entre elles. Un intérêt de cette approche réside en ce qu'elle permet une structuration non supervisée de succession de symboles en unités plus longues et de longueur variable, à partir de simples considérations statistiques. Par ailleurs, le modèle s'avère être un concurrent sérieux des n-grams, pour la modélisation de langage, puisqu'il permet d'obtenir une meilleure modélisation du corpus testé, avec un nombre d'unité moindre que les bigrams ou les trigrams.

On peut imaginer au moins 3 axes potentiels pour développer ce travail :

- au niveau acoustique ou phonétique, en utilisant le modèle de multigrams pour définir automatiquement des unités élémentaires pour la reconnaissance et la synthèse de parole,

- en tant que modèle de langage, en évaluant le modèle de multigrams au sein d'un système de reconnaissance de parole, par comparaison aux n-grams,

- en analyse et en compréhension de langage naturel, par exemple, en recherchant des équivalences sémantiques entre séquences de mots, et en étiquetant chaque classe par un concept spécifique.

REFERENCES BIBLIOGRAPHIQUES

[1] F. JELINEK : Self-organized language modeling for speech recognition, in Readings in Speech Recognition. pp. 450-506. *Ed. A. Waibel & K.F. Lee. Morgan Kaufmann Publishers Inc., San Mateo, Calif., 1990.*

[2] B.S. ATAL : Efficient coding of LPC parameters by temporal decomposition, proc. IEEE-ICASSP, pp. 81-84. *Boston, 1983.*

[3] F. BIMBOT, G. CHOLLET, P. DELEGLISE, C. MONTACIE : Temporal decomposition and acoustic-phonetic decoding of speech, proc. IEEE-ICASSP, pp. 445-448. *New-York, 1988.*

texte original :

```
[1]   blessed is the man that walketh not in the counsel of the ungodly nor standeth in the way
      of sinners nor sitteth in the seat of the scornful
[2]   but his delight is in the law of the lord and in his law doth he meditate day and night
[3]   and he shall be like a tree planted by the rivers of water that bringeth forth his fruit
      in his season his leaf also shall not wither and whatsoever he doeth shall prosper
[4]   the ungodly are not so but are like the chaff which the wind driveth away
[5]   therefore the ungodly shall not stand in the judgment nor sinners in the congregation of
      the righteous
[6]   for the lord knoweth the way of the righteous but the way of the ungodly shall perish
[7]   why do the heathen rage and the people imagine a vain thing
[8]   the kings of the earth set themselves and the rulers take counsel together against the
      lord and against his anointed saying
[9]   let us break their bands asunder and cast away their cords from us
[10]  he that sitteth in the heavens shall laugh the lord shall have them in derision
[11]  then shall he speak unto them in his wrath and vex them in his sore displeasure
```

texte en entrée :

```
[1]   blessedisthemanthatwalkethnotinthecounseloftheungodlynorstandethinthewayofsinnersnorsittet
      hintheseatofthescornful
[2]   buthisdelightisinthelawofthelordandinhislawdothhemeditatedayandnight
[3]   andheshallbelikeatreeplantedbytheriversofwaterthatbringethforthhisfruitinhisseasonhisleafa
      lsoshallnotwitherandwhatsoeverhedoethshallprosper
[4]   theungodlyarenotsobutarelikethechaffwhichthewinddrivethaway
[5]   thereforetheungodlyshallnotstandinthejudgmentnorsinnersinthecongregationoftherighteous
[6]   forthelordknoweththeweyoftherighteousbutthewayoftheungodlyshallperish
[7]   whydotheheathenrageandthepeopleimagineavainthing
[8]   thekingsoftheearthsetthemselvesandtherulerstakecounseltogetheragainstthelordandagainsthisa
      nointedsaying
[9]   letusbreaktheirbandsasunderandcastawaytheircordsfromus
[10]  hethatsittethintheheavensshalllaughthelordshallhavetheminderision
[11]  thenshallhespeakuntotheminhiswrathandvextheminhissoredispleasure
```

texte segmenté :

```
[1]   bless ed isthe man that walk eth not inthe couns el ofthe un godly n or stand eth inthe
      way of sin n ers n or sitte th inthe se at ofthe s corn ful
[2]   but his d e light is inthe law ofthe lord and inhis law do th he medit at e day and night
[3]   and he shall be likea t re e plant ed bythe river sof water that bring eth forth his fruit
      inhis s eason his l ea f also shall not with erand what so ever he doeth shall prosp er
[4]   the un godly are not so but are like the ch a f f which thew in d drive th away
[5]   ther efore the un godly shall not stand inthe judgm ent n or sin n ers inthe co ngreg
      ation ofthe right eous
[6]   for the lord know eth the way ofthe right eous but the way ofthe un godly shall peris h
[7]   why do the he athen ra g eand the peopl e i magin ea vain thing
[8]   the kings ofthe earth set the mselv esand the rule r s take couns e ltoge ther again stthe
      lord and again s this an o inted say ing
[9]   letus break their b and s as under and cast away their c ord s from us
[10]  he that sitte th inthe he avens shall laugh the lord shall have them i nder is ion
[11]  the n shall he speak unto them inhis wrath and ve x them inhis so red is pleas ure
```

Figure 1 : Psaumes du Vieux Testament, en anglais (King James version); 11 premiers versets.

De haut en bas : texte original, texte sans espace entre les mots, et texte segmenté par le modèle de multigrams (longueur max. : 5, pénalisation : 1.96). Les frontières de séquences sont représentées par des espaces.

```
( WHICH AIRLINES ) ( DEPART FROM <CITY> )
( FIND THE CHEAPEST ONE-WAY FARE ) ( FROM <CITY> TO <CITY> )
( FIND ) ( ALL ) ( FLIGHTS LEAVING <CITY> FOR <CITY> ) ( WHICH ) ( DEPART ) ( BEFORE <HOUR> IN THE MORNING )
( WHAT ) ( TYPE OF AIRCRAFT IS USED ) ( ON FLIGHT <NUMBER> )
( FIND ) ( FLIGHTS ON <AIRLINE> AIRLINES ) ( FROM <CITY> TO <CITY> ) ( THAT ) ( STOPOVER IN <CITY> )
( ARE THERE ANY <AIRLINE> FLIGHTS ) ( FROM <CITY> TO <CITY> ) ( STOPPING IN <CITY> )
( PLEASE ) ( DISPLAY ) ( AN ) ( ITINERARY ) ( FOR <AIRLINE> FLIGHT <NUMBER> ) ( FROM <CITY> TO <CITY> )
( WHAT TIME DOES FLIGHT <NUMBER> ) ( STOP IN <CITY> )
( WHAT TIME DOES FLIGHT <NUMBER> ) ( ARRIVE IN <CITY> )
( WHAT IS THE EARLIEST FLIGHT ) ( FROM <CITY> TO <CITY> )
```

Figure 2 : Texte de la base de données ATIS ; 10 premières phrases de test.

Cet exemple présente la segmentation obtenue en modélisation de langage par multigrams (longueur max. : 5, pénalisation : 1.96). Les frontières de séquences sont représentées ici par des parenthèses.

Interaction entre le décodage acoustico-phonétique et l'analyse lexicale grâce aux techniques du raisonnement hypothétique : le système DAPHNÉ

Sylvie COSTE-MARQUIS

CRIN/CNRS & INRIA-Lorraine
BP 239 54506 Vandœuvre-lès-Nancy

Résumé

Nous présentons dans cet article un système à bases de connaissances pour la reconnaissance automatique de la parole continue multilocuteur, le système DAPHNÉ, que nous avons testé sur des phrases constituées d'occlusives et de voyelles. Nous montrons comment le raisonnement hypothétique est utile, à la fois pour émettre des hypothèses expliquant les observations acoustiques, pour les gérer, et pour réviser les données quand une incohérence apparaît. Nous proposons un modèle d'interaction entre le décodage acoustico-phonétique et l'analyse lexicale qui permet d'éviter d'explorer des solutions reposant sur des hypothèses incohérentes. Un exemple illustrant notre approche est développé dans l'article.

Mots-clés : décodage acoustico-phonétique, raisonnement hypothétique, occlusive, interaction lexicale.

1 Introduction

Notre but est de construire un système de reconnaissance de la parole continue multilocuteur qui utilise des connaissances phonétiques. Notre méthode repose sur l'extraction d'indices pertinents du signal de parole interprétés grâce à des connaissances phonétiques élaborées en collaboration avec une phonéticienne, A. Bonneau. Nous utilisons les détecteurs acoustiques automatiques implantés par Y. Laprie dans l'outil Snorri [7]. Dans ce cadre, nous proposons le système DAPHNÉ, que nous avons testé sur des phrases formées d'occlusives et de voyelles.

Pour résoudre les problèmes de la reconnaissance de parole continue multilocuteur qui concernent notamment l'incertitude des données, l'absence éventuelle de certains indices acoustiques, la dépendance du pouvoir de discrimination d'un indice acoustique par rapport à sa réalisation, un système de reconnaissance automatique de la parole continue doit pouvoir engendrer, gérer et réviser des hypothèses. Dans cet objectif, une approche logique permet d'utiliser de nombreux travaux en intelligence artificielle [13]. De plus, cette approche permet une meilleure compréhension à la fois de la connaissance et du savoir-faire de l'expert ; en particulier, les inférences valides (déduction) sont différenciées de la génération d'hypothèses (abduction).

Dans notre système, les incohérences rencontrées lors du décodage acoustico-phonétique sont gérées par une interaction avec les détecteurs automatiques d'indices acoustiques ou avec l'étape de segmentation. L'interaction avec l'analyse lexicale permet également de pallier l'absence de certains indices acoustiques. Enfin, cette dernière interaction permet de distinguer les informations fiables fournies par le décodage acoustico-phonétique de celles à remettre en cause ; elle permet donc à l'analyse lexicale de se focaliser sous les hypothèses intéressantes.

Notre approche est analogue à celle des travaux menés à l'Université de l'Ohio [8]. Dans ces travaux, la reconnaissance de la parole est considérée comme un processus abductif à plusieurs niveaux [10]. Les données d'un niveau sont les hypothèses-solutions proposées par le niveau inférieur. R. Fox [8] propose un système capable de reconnaître un petit vocabulaire à partir de gestes articulatoires eux-mêmes inférés à partir de mouvements de cibles détectés par rayons X.

Notre article est organisé comme suit. Dans une première partie, nous exposons notre méthode de décodage acoustico-phonétique des occlusives et voyelles orales. Nous décrivons ensuite l'étape de révision des hypothèses. Enfin, nous présentons l'interaction avec le niveau lexical. Nous terminons par une description du système et quelques résultats avant de conclure et discuter les perspectives de notre travail.

2 Décodage acoustico-phonétique

Dans un premier temps, nous avons mis en œuvre notre modèle sur le décodage des occlusives en contexte vocalique. Nous présentons brièvement le principe du décodage des occlusives. Nous décrivons ensuite l'identification des voyelles et le processus global de reconnaissance des VCV.

2.1 Identification des occlusives

Le processus d'identification des occlusives et les indices acoustiques sont respectivement décrits dans [1, 3]. Les principaux indices des occlusives sont fournis par :

- les trajectoires formantiques à la frontière avec les voyelles adjacentes,
- le bruit d'explosion (nous utilisons uniquement les caractéristiques de la barre d'explosion).

Nos indices et leur stratégie d'utilisation dépendent des voyelles du contexte. Nous avons distingué trois classes de voyelles : les voyelles d'arrière, les voyelles antérieures et les voyelles centrales.
Nous avons défini deux sortes d'indices. Les indices «forts» sont à la fois très prononcés et très discriminants. S'ils sont observés, ils permettent soit l'identification directe de l'occlusive (indices «forts» de préférence), soit l'élimination directe de certains candidats (indices «forts» d'exclusion). Si de tels indices ne sont pas observés, l'identification de l'occlusive repose sur des indices moins prononcés ou moins discriminants (appelés indices «faibles»).
Le schéma de raisonnement à partir de ces indices est abductif/déductif [12].

2.2 Identification des voyelles

Dans un premier temps, nous nous sommes limitée aux voyelles orales. Pour la reconnaissance des voyelles, nous avons utilisé les connaissances et le savoir-faire du système APHODEX [4], dont nous rappelons ici le principe de fonctionnement. Les fréquences des trois premiers formants sont comparées à des valeurs cibles, définies pour chaque formant de chaque voyelle. Cette comparaison utilise des écarts admissibles, définis pour chaque formant de chaque voyelle en fonction du lieu d'articulation du son précédent. Nous utilisons le suivi de formants [11] pour trouver les fréquences des trois premiers formants.

2.3 Stratégie générale

L'ensemble du processus de reconnaissance est illustré sur la figure 1. Les voyelles sont d'abord identifiées (1). Chaque hypothèse vocalique dépend du contexte gauche. À partir des hypothèses vocaliques, différentes classes de voyelles sont proposées. Plusieurs classes sont possibles si la réalisation d'une voyelle se trouve à la frontière entre plusieurs classes. À l'aide des hypothèses sur les classes de voyelles, les indices acoustiques des occlusives adjacentes sont interprétés (2) et des solutions proposées (3). Des hypothèses sont ensuite émises pour les VCV en tenant compte de la dépendance réciproque des hypothèses de voyelles et d'occlusives (4). Si une incohérence empêche DAPHNÉ de trouver une solution, l'étape de remise en cause (5) essaie de la lever.

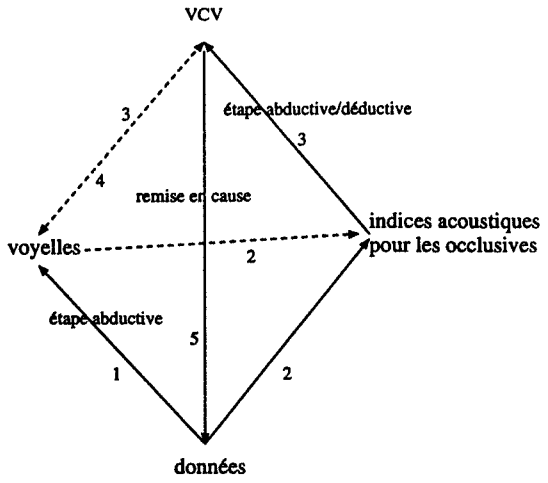


FIG. 1 - *Processus de décodage des VCV.*

2.4 Exemple

Nous décrivons le fonctionnement de notre système sur un exemple. La phrase choisie est tirée de notre corpus. Son spectrogramme est donné sur la figure 2, avec les variations du pitch. Nous nous intéressons au décodage des sons 8, 9 et 10.
Les hypothèses pour la première voyelle (8) sont /ɔ/ ou /œ/ ou /ø/ [1]. Pour la deuxième voyelle (10) /ɔ/ ou /œ/ sont proposés. Ce cas est particulièrement défavorable, car nous avons six contextes possibles (trois à gauche × deux à droite) à explorer pour identifier l'occlusive.
Dans ces six contextes possibles, les deux indices forts suivants sont observés :

- barre d'explosion dentale (BD) (diffuse, avec un maximum d'énergie en haute fréquence)
- transition de F2 excluant une dentale à gauche (TD) (elle ne prend pas du tout la direction du locus, entre 1500 Hz et 2000 Hz).

Les deux indices forts observés sont contradictoires et empêchent l'identification de l'occlusive (9). La détection d'une telle contradiction appelle une révision des données acoustiques qui l'ont provoquée.

3 Gestion des incohérences

Notre méthode assure la cohérence du raisonnement tout au long du processus de reconnaissance. Quand une incohérence est rencontrée, deux cas sont possibles. Dans le premier cas, l'incohérence est présente seulement dans certains contextes vocaliques

1. Le fait d'obtenir comme solution trois voyelles appartenant à trois classes différentes est rare mais justifié ici. En effet, dans le plan F1×F2, la réalisation de la voyelle (8) se situe au centre du triangle vocalique. Chacune de nos classes correspondant à un angle du triangle, la voyelle (8) peut appartenir à chacune des classes.

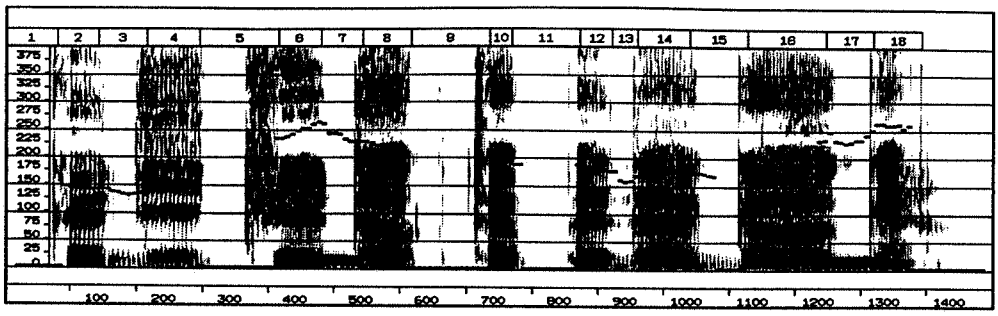


FIG. 2 - *Spectrogramme de la phrase «Que dis-tu, docte pédagogue?».*

détectés. Nous supposons alors que ces contextes sont incohérents et ils sont éliminés. Le processus de reconnaissance se poursuit avec les autres contextes. Dans le cas où cette incohérence élimine tous les contextes possibles, nous remettons en cause les données qui l'ont provoquée dans chacun des contextes. Cette étape de révision est aussi utilisée quand aucun indice n'a été trouvé.
La contradiction peut provenir d'une erreur d'un détecteur automatique, de l'étape de segmentation ou d'un effet de coarticulation à moyen terme. Dans la suite, nous exposons le mécanisme de remise en cause que nous avons implanté dans notre système DAPHNÉ.

3.1 Révision des données

Nous proposons deux mécanismes de remise en cause. Le premier concerne une éventuelle erreur d'un détecteur acoustique, le deuxième permet de corriger une erreur de l'étape de segmentation.
La première méthode de révision proposée permet de vérifier les valeurs des formants proposées par le suivi automatique. Nous utilisons les racines de la LPC (ou les maxima du spectre) dans la partie stable de la voyelle et nous vérifions que les valeurs trouvées peuvent correspondre à la réalisation d'une voyelle. Si ces valeurs sont différentes de celles proposées par le suivi de formants, nous révisons les données initiales fournies par le suivi de formants. Cette révision a au moins deux effets.

- De nouvelles hypothèses de voyelles vont entraîner de nouvelles hypothèses de classes, ce qui peut conduire à une interprétation différente des indices acoustiques des occlusives.

- Des indices situés sur les transitions vont disparaître.

Le deuxième mécanisme met en cause l'étape de segmentation dans certains cas précis. La segmentation est révisée en fonction de la durée de l'occlusive grâce aux données de l'étude de M. Wajskop [15]. Si les deux indices forts contradictoires sont de part et d'autre de l'occlusive (par exemple, les transitions à gauche «contredisent» la barre d'explosion), le système suppose que le son est une suite de deux occlusives ou une pause (éventuellement suivie d'une occlusive), selon la durée du segment. Si les indices contradictoires sont à droite de l'occlusive (les transitions à droite «contredisent» la barre d'explosion), le système suppose que l'occlusive sourde proposée par la segmentation peut être une pause si la durée du segment le permet.
Dans ce deuxième mécanisme, les indices et leur interprétation ne changent pas, mais la contradiction peut être levée et une solution proposée. Dans le cas d'une double occlusive, l'identification de la première occlusive repose uniquement sur les transitions à gauche (avec la voyelle précédente).
Les deux mécanismes sont utilisés en parallèle. Cependant, le mécanisme de révision des formants est prioritaire. Les faits mis en cause sont enlevés de la base de faits et remplacés par d'autres, qui sont déclarés incohérents avec les premiers. Le mécanisme de raisonnement est relancé sur les nouveaux faits. Pour compléter le mécanisme de gestion des incohérences actuel, il faudra étudier les effets de coarticulation à moyen terme.
Bien sûr, une solution cohérente peut être fausse. C'est seulement l'interaction avec un niveau lexical qui permettra peut-être de détecter une erreur. Si une occlusive ne peut pas être reconnue par le décodage acoustico-phonétique, même après la remise en cause, aucune hypothèse n'est émise. Le système laisse le niveau lexical proposer des solutions.

3.2 Exemple (suite)

Considérons à nouveau notre exemple. DAPHNÉ ne remet pas en cause le suivi de formants (ce qui est justifié). En revanche, la durée de l'occlusive et la nature de l'opposition entre les indices permet de supposer que le son à reconnaître est un groupe de deux occlusives sourdes. Sous cette hypothèse, le système propose une nouvelle segmentation et insère une occlusive (9bis). À partir de cette segmentation, DAPHNÉ propose la suite /p t/ dans les six contextes vocaliques proposés. La reconnaissance du son (9) repose uniquement sur la transition de F2 à gauche.

Bien que peu fréquent, cet exemple réel de double occlusive illustre bien l'intérêt de notre approche.

4 Interaction avec un niveau lexical

Dans cette partie, nous proposons une interaction forte entre le décodage et le niveau lexical. Nous présentons d'abord le codage du lexique et le mécanisme d'accès lexical. Ensuite, nous montrons comment l'analyse lexicale utilise le décodage acoustico-phonétique pour disposer de points d'ancrage et pour éviter d'explorer des espaces de recherche reposant sur des hypothèses incohérentes entre elles.

4.1 Lexique

Nous avons enregistré un corpus de 22 phrases composées de voyelles et d'occlusives. Notre lexique comprend les mots de ce corpus, soit 94 mots. Pour chaque mot, nous avons codé les différentes réalisations possibles, en indiquant éventuellement les réalisations préférées. Nous avons également codé les liaisons possibles.
La recherche des mots dans le lexique se fait sur la base d'un son et son contexte. Pour diminuer le nombre de mots candidats, nous utilisons un indice prosodique très simple et bien adapté à notre type de phrases, lui-même très simple. Si une montée importante de F0 accompagnée d'un allongement de la durée d'une voyelle est détectée, le système suppose qu'il y a une fin de mot. Deux cas se présentent : la voyelle suivante est un /ə/ (un mot se termine sur la syllabe suivante) ou n'en est pas un (un mot se termine sur la syllabe où l'indice est observé). Les deux hypothèses mutuellement exclusives sont émises de manière indépendantes.

4.2 Interaction

Les hypothèses de mots émises à partir d'indices «forts» de préférence constituent des informations fiables dans la phrase. Les hypothèses de mots fournies par l'accès lexical et les hypothèses de sons sont mises en correspondance en utilisant des contraintes locales. L'analyse lexicale consiste à expliquer par une suite de mots les hypothèses fournies par le décodage. La prononciation de chaque mot induit la réalisation d'une suite de sons. À partir des sons proposés par le décodage, l'analyse lexicale essaie d'inférer de manière abductive des explications, chaque explication étant une suite de mots.
Des règles autorisent la mise en correspondance malgré des différences, d'autres interdisent certaines mises en correspondance.
Les indices «forts» du décodage forment des points d'ancrage pour le niveau lexical. Les indices «forts» de préférence doivent donc être respectés et les indices «forts» d'exclusion ne doivent pas être contredits. De plus, comme les indices acoustiques des occlusives dépendent des classes du contexte vocalique, les voyelles ne peuvent pas être substituées par des voyelles appartenant à d'autres classes sans nouvelle interprétation de ces indices. Deux de ces règles sont :

- Une substitution entre voyelles proches mais de classes différentes est permise si l'interprétation des indices «forts» des occlusives adjacentes ne change pas.
- La substitution entre voyelles de classes différentes (/i/ et /u/, par exemple) n'est pas permise. Si elle est nécessaire pour trouver une solution, le mécanisme de remise en cause décrit précédemment est utilisé.

D'autres règles sont essentiellement de bon sens et ne sont pas décrites ici. Toutes ces règles sont utiles pour trouver une solution dans les cas où le décodage n'a pas trouvé une suite de sons qui serait exactement en correspondance avec une suite de mots.

4.3 Exemple (fin)

Nous illustrons maintenant cette interaction sur notre exemple. L'indice prosodique décrit précédemment est observé sur la voyelle (6). Deux hypothèses sont alors émises en parallèle :
(1) un mot se termine sur la voyelle (6),
(2) un mot se termine par un /ə/ sur la voyelle (8).
L'analyse lexicale explore les deux possibilités. Nous nous intéressons au début de la deuxième partie de la phrase (après l'hypothèse de fin de mot). Le son (7) a été reconnu comme /d/.

Sous l'hypothèse (1)
L'accès lexical propose des mots pour expliquer chaque son en contexte. Aucun des mots proposés n'est en correspondance complète avec les hypothèses fournies par le décodage acoustico-phonétique.
Les indices forts constituent les informations fiables du décodage acoustico-phonétique. Par conséquent, nous devons expliquer l'indice fort (BD) et respecter les déductions provoquées par l'indice fort d'exclusion (TD). Seuls les mots contenant les groupes d'occlusives /p t/ et /k t/ vérifient ces deux conditions, le groupe /p t/ étant celui préféré mais non imposé par le décodage acoustico-phonétique. Le décodage acoustico-phonétique n'a pas suggéré /k/ comme solution pour le son (9) à cause d'un effet de coarticulation à moyen terme du /d/ sur la fin du /ɔ/ qui affecte à cet endroit la transition de F2. Les mots de notre lexique correspondants sont : «adopta» et «docte». «adopta» est contradictoire avec l'hypothèse (1) et est donc éliminé par DAPHNÉ. «docte» est retenu comme solution.

Sous l'hypothèse (2)
Nous ne détaillerons pas le processus de recherche mais aucune solution n'est trouvée sous cette hypothèse.

Seul le mot «docte» est donc proposé par DAPHNÉ pour expliquer les sons (6 7 8 9 9bis 10).

5 Le système DAPHNÉ

5.1 Présentation

La méthode de reconnaissance présentée dans cet article a été implantée dans le système DAPHNÉ. Les données acoustiques du système sont des événements acoustiques fournis pas Snorri [7], le suivi de formant utilisé est celui d'Y. Laprie [11], la décomposition de l'explosion en attaque et bruit de friction, puis en pics d'énergie est décrite dans [3].
DAPHNÉ est construit à partir de l'outil de développement X-TRA [5]. Les détails d'implantation de l'étape abductive/déductive du décodage sont donnés dans [2].
Le système fonctionne complètement pour les occlusives suivies de voyelles d'arrière. Pour les autres classes de voyelles, la reconnaissance des occlusives repose sur les indices liés aux transitions formantiques. Les indices situés sur la barre d'explosion pour les occlusives suivies de voyelles centrales sont actuellement en cours de développement.

5.2 Résultats

Les tests ont été effectués pour des occlusives suivies de voyelles d'arrière. Nous donnons d'abord des résultats sur les indices de la barre d'explosion. Nous décrivons ensuite les résultats obtenus sur la reconnaissance des VCV.

5.2.1 Méthodologie

Nous avons utilisé deux corpus de parole continue. Chaque corpus est composé de phrases simples. Le premier corpus [14] est constitué de phrases prononcées par 16 locuteurs. Dans le but de tester notre système, nous avons enregistré un autre corpus de 22 phrases composées uniquement d'occlusives et de voyelles, chaque phrase a été prononcée trois fois par 4 locuteurs masculins.
Nous avons extrait les répétitions de 10 locuteurs du premier corpus et une répétition de notre corpus pour ajuster les seuils de nos indices. Pour faire nos tests, nous avons utilisé les répétitions des 6 autres locuteurs du premier corpus (corpus1) et la deuxième répétition de notre corpus (corpus2). La segmentation des VCV est manuelle, par contre les événements acoustiques sont extraits de manière automatique.
Nous avons effectué une première série de tests uniquement sur les indices situés sur la barre d'explosion, en supposant que la voyelle suivante était correctement reconnue comme une voyelle d'arrière. Dans une deuxième série de tests, nous avons testé le système complet de décodage, c'est à dire la reconnaissance complète des VCV comprenant la reconnaissance des voyelles et des occlusives, cette dernière utilisant la barre d'explosion et les transitions. Dans cette série, les voyelles sont donc également proposées par notre système.

5.2.2 Résultats

Le tableau 1 résume les résultats des indices situés sur la barre d'explosion pour les deux corpus.
Aucune fausse alarme pour les indices «forts» n'a été détectée. Nous n'avons pas défini d'indice fort pour /p/ suivi par une voyelle d'arrière. Nous remarquons que les résultats sont moins élevés pour /p/ dans le premier corpus. D'une part, la plupart des /VpV/ ont été extraits de mots grammaticaux. D'autre part, nous verrons que les transitions améliorent les résultats. Certains /k/ sont suivis de /ɔ/ très centralisés et ont plutôt une barre d'explosion caractéristique de /k/ suivis par une voyelle centrale (élévation de la fréquence et augmentation de la largeur de bande du pic principal).

TAB. 1 - *Indices (barre d'explosion)*

Corpus	Indices forts (déclenchement)	Bonne solution (indices forts + faibles)
corpus1		
p (89)	–	68%
t (92)	47%	88%
k (105)	28%	89%
corpus2		
p (25)	–	80%
t (55)	47%	93%
k (32)	34%	97%

Le tableau 2 donne les résultats de reconnaissance des occlusives dans les VCV. Nous remarquons que la reconnaissance est améliorée par l'emploi des transitions, en particulier pour /p/. Un /t/ est proposé à la place d'un /p/ à cause d'un effet de coarticulation à moyen terme sur la transition de F2. Dans 8% des cas, DAPHNÉ n'a proposé aucune solution, principalement à cause de problèmes de reconnaissance des voyelles adjacentes. Les erreurs sur le contexte vocalique sont aussi la cause principale des 4% d'erreur, en particulier pour les /k/. Notons encore qu'il n'y a aucune fausse alarme d'indices «forts».

TAB. 2 - *Reconnaissance VCV (barre d'explosion + transitions + voyelle)*

Corpus	Bonne solution
p (25)	88%
t (25)	92%
k (25)	76%

Finalement, le système a reconnu correctement 85% des VCV et dans 3% des cas, les solutions correctes ont été proposées en deuxième solution. Dans la plupart des autres cas, le problème est dû à des erreurs de reconnaissance sur les voyelles. Des expériences plus complètes sur des phrases entières sont en cours de façon à évaluer DAPHNÉ sur une plus grande échelle.
Le système sera complété dans les directions suivantes. Une extension de notre système consiste à enrichir la base de connaissances. Une première direction est la recherche d'un autre indice d'identification pour le /p/ (par exemple, l'amplitude de la barre d'explosion [16] [6]). Il est aussi nécessaire d'étudier les effets de coarticulation à moyen terme. Enfin, il sera nécessaire d'étendre DAPHNÉ aux autres consonnes du français.

6 Perspectives et conclusion

Nous avons présenté un système de reconnaissance des phrases constituées d'occlusives et de voyelles. Ce système tire profit des techniques de raisonnement hypothétique qui nous permettent d'émettre, gérer et réviser de nombreuses hypothèses à plusieurs niveaux de la reconnaissance. Un point fort de notre approche réside dans la remise en cause des données provenant d'un niveau inférieur (par exemple, événements acoustiques fournis par des détecteurs automatiques pour le décodage acoustico-phonétique) quand une incohérence est détectée. Notre approche de DAP nous a également conduit à un modèle d'analyse lexicale fondé sur les indices forts du décodage acoustico-phonétique. Cette utilisation des informations fiables du décodage nous permet de guider l'analyse lexicale et d'élaguer l'espace de recherche.
Une première perspective est l'intégration d'un module phonologique qui permettra de gérer les phénomènes d'altération entre les mots et à l'intérieur des mots. Enfin, une perspective à plus long terme consiste à étendre notre modèle de raisonnement aux plus hauts niveaux de la reconnaissance [9], dans un premier temps à l'analyse syntaxique.

Références

[1] A. Bonneau, F. Charpillet, S. Coste, J.-P. Haton, Y. Laprie, and P. Marquis. A Model for Hypothetical Reasoning applied to Speech Recognition. In *Proc. of the 10th European Conference on Artificial Intelligence*, pages 583–587, Vienna, Austria, 1992.

[2] A. Bonneau, F. Charpillet, S. Coste, J.-P. Haton, Y. Laprie, and P. Marquis. An Overview of some Applications Using Assumption-based reasoning. In *Proc. of the ECAI workshop on Application of Reason Maintenance Systems*, Vienna, August 4th, 1992.

[3] A. Bonneau, S. Coste, L. Djezzar, and Y. Laprie. Two levels acoustic cues for consistent stop identification. In *Proc. of the International Conference on Spoken Language Processing*, volume 1, pages 511–514, Banff (Alberta, Canada), 1992.

[4] N. Carbonell, J.-P. Damestoy, D. Fohr, J.-P. Haton, and F. Longchamp. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. In *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 1201–1204, Tokyo, Japan, 1986.

[5] F. Charpillet, Ph. Théret, and J.P. Haton. X-TRA: un moteur d'inférence comportant deux modes de compilation des règles TREAT et RETE et un système de maintien de vérité de type ATMS. In *Actes des journées internationales sur les systèmes experts*, Avignon, France, 1988.

[6] T. Edwards. Multiple features analysis of intervocalic English plosives. *Journal of the Acoustical Society of America*, 69(2):535–547, 1981.

[7] D. Fohr and Y. Laprie. Snorri: An Interactive Tool for Speech Analysis. In *Proc. of the 1st European Conference on Speech Communication and Technology*, pages 669–672, Paris (France), 1989.

[8] R. Fox. *Layered Abduction for Speech Recognition from Articulation.* PhD thesis, The Ohio State University, Columbus, Ohio, 1992.

[9] J. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63:69–142, 1993.

[10] John R. Josephson and Susan G. Josephson, editors. *Abductive Inference: Computation, Philosophy, Technology.* Forthcoming, Cambridge University Press, 1994.

[11] Y. Laprie and M.-O. Berger. Active Models for Regularizing Formant Trajectories. In *Proc. of the International Conference on Spoken Language Processing*, volume 1, pages 815–818, Banff (Alberta, Canada), 1992.

[12] P. Marquis. Towards Data Interpretation by Deduction and Abduction. In *Proc. AAAI Workshop on Abduction*, pages 59–63, Anaheim, California, 1991.

[13] J. Martins. The Truth, the Whole Truth, and Nothing But the Truth: An Indexed Bibliography to the Literature of Truth Maintenance Systems. *AI Magazine (special issue)*, pages 7–25, 1990.

[14] O. Mella and M.-C. Haton. Méthodologie d'étude de la pertinence de paramètres phonétiques et acoustiques pour la reconnaissance du locuteur. In *Actes du séminaire sur la variabilité et la spécificité des locuteurs, GRECO-PRC Communication Homme-Machine*, Marseille Luminy, 1989.

[15] M. Wajskop. Segmental Durations of French Intervocalic Plosives. In B. Lindblom and S. Öhman, editors, *Frontiers of Speech Communication Research*, pages 109–134. Academic Press, London, 1979.

[16] V. Zue. *Acoustic Characteristics of Stop Consonants: A Controlled Study.* PhD thesis, MIT, 1976.

DÉTECTION DE MOTS CLÉS DANS UN DISCOURS CONTINU

BÉCHET F., IGOUNET S., MÉLONI H., GILLES P.

LABORATOIRE D'INFORMATIQUE - FACULTÉ DES SCIENCES
33, RUE LOUIS PASTEUR - 84000 AVIGNON - FRANCE

RÉSUMÉ

Dans le cadre de la reconnaissance automatique de la parole à l'aide de techniques analytiques, nous proposons un système de détection de mots cibles dans de la parole continue (word spotting). Ces travaux présentent un ensemble de modules utilisant tous une représentation explicite de connaissances phonétiques et lexicales. Le système se compose de trois niveaux : un niveau de décodage acoustico-phonétique ascendant, un module de sélection de mots clés probables et enfin un niveau de vérification contextuelle descendante des hypothèses proposées. Les premiers tests effectués visent à évaluer la qualité de notre module de rejet contextuel d'hypothèses lexicales.

1. INTRODUCTION

La détection de mots clés ou *word spotting* consiste à analyser une phrase prononcée par un locuteur afin d'en isoler les items connus du lexique. La grande différence par rapport à un système de reconnaissance de parole continue classique est la part réduite des informations connues par rapport aux données analysées. La reconnaissance par *word spotting* n'a aucune information sur la zone où les mots cibles ont été réalisés. Elle doit donc résoudre deux problèmes : quels items du lexique ont été prononcés et à quels endroits dans le signal de parole ? Cette méthode peut donc se comparer à une reconnaissance de mots dans un milieu bruité où le bruit se compose des phénomènes hors-langage et des mots inconnus du système.
Les principaux domaines d'application de cette technique sont l'interrogation de bases de données, le réveil d'un système mis à l'état de veille ou encore la commande vocale en environnement bruité. L'utilisateur a ainsi la latitude d'exprimer sa requête de façons très diverses pourvu qu'il prononce le ou les mots-clés requis. Il n'est donc pas nécessaire de concevoir une grammaire générale des formules d'interrogation possibles, ni d'utiliser des vocabulaires importants. C'est le contexte d'utilisation, couplé à une gestion rigoureuse du dialogue (choix limité, prise en compte de l'historique des requêtes, etc.) qui permet de limiter à chaque étape de l'interrogation l'ensemble des mots clés nécessaires pour répondre à la demande de l'utilisateur.
Les nombreuses études actuelles sur ce sujet [Rose, 90, 93; Weintraub, 93; Wilpon, 90; Boite 93; Rohlicek, 93; Lleida, 93] utilisant des méthodes statistiques à base de modèles de Markov, se sont attachées au problème de la représentation des informations inconnues. En effet comment doit-on définir de manière statistique des phénomènes dont on ne sait rien a priori ? Les principales méthodes employées pour résoudre ce paradoxe ont été l'utilisation de grands lexiques cherchant une reconnaissance optimale des items prononcés afin d'en extraire les mots-clés et la définition d'un "modèle poubelle" censé représenter au mieux les données inconnues "polluant" le signal d'entrée.
Notre démarche ne cherche pas à représenter les informations hors mots-clés. L'apprentissage n'étant réalisé que sur les parties stables des phonèmes, notre processus de reconnaissance est beaucoup moins perturbé par l'apparition de nouveaux mots.

2. MÉTHODE GÉNÉRALE

Les méthodes présentées de type analytique sont fondée sur la représentation explicite de connaissances phonétiques et lexicales. Le système est dépendant du locuteur mais l'apprentissage nécessaire à l'intégration d'un nouvel utilisateur est très rapide. Il est constitué de la prononciation de 34 mots représentant l'ensemble des phonèmes du français dans plusieurs contextes peu déformants. La prise en compte de la variabilité inter-locuteurs se fait par l'utilisation, pour chaque individu, d'une référence spectrale associé à chaque phonème. Le but ici n'est pas de décrire précisément les sons - qui de toute manière sont plus ou moins fortement modifiés par le contexte - mais de proposer une forme moyenne proche de la réalisation idéale du phonème par un locuteur. Le système se compose de trois niveaux :

- un niveau acoustico-phonétique proposant un décodage phonétique du signal traité,
- un niveau d'accès lexical qui permet d'émettre des hypothèses sur la localisation et l'identification des mots-clés recherchés [Béchet, 94],
- un niveau de rejet d'hypothèses qui épure, grâce à des processus de filtrage phonétique contextuels, la cohorte d'items proposée par l'accès lexical.

Nous allons détailler maintenant ces trois niveaux avant d'évoquer le protocole de test qui a permis d'évaluer nos modules.

3. NIVEAU PHONÉTIQUE

Ce niveau est constitué, dans notre système, par un module de Décodage Acoustico-Phonétique (DAP) [Méloni, 91; Gilles, 93].
L'unité phonétique utilisée par ce module est le phonème. Celui-ci semble s'imposer dans les systèmes demandant un apprentissage rapide avec une forte indépendance vis-à-vis du corpus où il est effectué. Les problèmes inhérents à l'emploi de cette unité nous ont

conduit à considérer deux phases distinctes dans notre processus de décodage phonétique :

- Une phase ascendante d'accès aux unités phonologiques qui privilégie un décodage fondé sur la localisation et l'identification des macro-classes et des macro-traits phonétiques et la "ressemblance" acoustique ou perceptive de certaines zones de signal avec les références des sons. Le résultat de cette phase se présente sous la forme d'un treillis d'unités phonétiques valuées par une mesure de la distance spectrale entre la référence stockée du phonème proposé et la portion de signal où le son a été réalisé.
- Une phase descendante qui va favoriser la recherche d'indices et de propriétés sur des paramètres et dans des intervalles de signal spécifiquement adaptés au contexte. C'est cette étape descendante qui nous permettra de filtrer la cohorte résultat.

4. NIVEAU LEXICAL

4.1. CODAGE PHONÉTIQUE

À l'issue du processus de DAP ascendant, nous disposons d'un treillis phonétique caractérisant le signal traité selon les spécifications suivantes [Gilles, 93] :

- pour un phonème localisé le DAP propose un ensemble de phonèmes valués,
- chaque phonème voyelle réalisé doit être présent dans le treillis,
- seuls les groupes consonantiques inter-vocaliques doivent être identifiés, à l'exception de certaines consonnes sonnantes placées dans un contexte de voyelles fermées (ex : le phonème /l/ dans la séquence /i l i/),
- dans un groupe consonantique réalisé identifiable, seul le son le plus fermé est obligatoire dans le treillis (ex : /t/ dans la séquence /e t r w a/).

Les règles concernant les suppressions des consonnes lors du décodage ainsi que le nombre important d'insertions pour chaque son réalisé ne permettent pas de rechercher directement dans le treillis, de manière efficace, les chaînes phonétiques des mots du lexique. Nous avons ainsi adopté une représentation intermédiaire, en macro-classes, entre le treillis et la décomposition phonétique de chaque mot. Cette représentation est basée sur une segmentation du treillis résultat en différentes zones appelées *familles* de phonèmes.

L'idée de base de cette représentation est de regrouper dans un ensemble appelé *famille* tous les phonèmes proposés par le DAP lors de la réalisation d'un son. Cette segmentation du treillis permet de traiter le problème de l'insertion importante d'hypothèses phonétiques lors du décodage en proposant une structure du mot prononcé. Nous discernons deux types de familles : voyelles et consonnes. Ainsi une famille représente l'ensemble des unités de même type (voyelle ou consonne) proposées par le DAP dont les noyaux stables – déterminés lors de la localisation des phonèmes – se chevauchent.

Une fois la segmentation en familles réalisée, le système calcule quelques paramètres supplémentaires afin de caractériser chacune des familles voyelles selon trois étiquettes :

- Voyelle réalisée (codée VO) : famille couvrant une zone sur laquelle une voyelle a été réalisée de façon certaine.
- Appui vocalique (codée EE) : famille représentant un schwa qui n'a pas été pleinement réalisé ou une épenthèse dans un groupe consonantique mais en aucun cas une voyelle réalisée.
- Voyelle incertaine (codée VX) : famille ne pouvant être classée avec certitude dans l'une des deux catégories précédentes.

Les familles consonnes seront étiquetées CO.

A partir de la décomposition en familles étiquetées du treillis, nous allons générer un ensemble d'hypothèses (appelées *images*) sur la structure vocalique/consonantique du mot prononcé. Ces images sont obtenues par l'application de trois règles de dérivation sur la configuration des familles trouvées sur le signal et serviront de lien d'accès aux données lexicales. Il faut noter que ces règles correspondent aux caractéristiques du décodage de notre module phonétique et qu'elles se composent. Voici l'énoncé de ces règles :

- Règle 1 : les familles étiquetées EE sont ignorées lors de la dérivation car elles ne correspondent pas à la réalisation effective d'un son voyelle. En effet, un appui vocalique entre deux phonèmes ou la réalisation incomplète d'un /ə/ peuvent induire l'insertion de phonèmes dans le treillis.
- Règle 2 : les familles étiquetées VX vont être interprétées soit comme des VO soit comme des EE, et donc engendrer deux modèles.
- Règle 3 : lorsque deux familles VO sont consécutives, elles peuvent dériver en une seule famille VO. Là encore, deux images sont engendrées. Cette règle dérive de l'objectif du DAP de localiser l'ensemble des noyaux vocaliques sur un signal donné. Certaines voyelles, surtout pour un débit de parole lent, peuvent ainsi être segmentés en plusieurs noyaux vocaliques distincts.

Le résultat de cette phase de génération est un ensemble d'*images*. Chaque *image* représente une interprétation possible du treillis phonétique concernant la structure du mot prononcé (figure 1).

Nous allons décrire maintenant le codage des données phonologiques avant d'examiner les processus d'accès qui permettent de réaliser ce lien.

4.2. CODAGE DES DONNÉES LEXICALES

Dans notre système les mots seront décrits par leurs décompositions en phonèmes. Le dictionnaire de mots-clés contient, pour chaque item, toutes les différentes variations phonologiques qui peuvent intervenir au niveau des liaisons inter-lexicales. Chaque variante fait l'objet d'une nouvelle entrée dans notre lexique.

Au niveau du codage de chaque item, deux règles de variation phonologique sont prises en compte. D'une part l'expression de certaines confusions vocaliques liées à l'usage et n'ayant pas d'effet trop négatif sur l'accès lexical : <e/ɛ>, <o/ɔ>, <ø/œ>, <ɛ̃,œ̃>. D'autre part, la

règle d'élision du schwa (/ə/) car elle remet en cause la structure des mots recherchés.
Le système va déduire un certain nombre de modèles pour chaque item du dictionnaire. Ces modèles représentent les *images* vocalique/consonantique décrites précédemment. Ils sont obtenus automatiquement par dérivation de deux sortes de règles: d'une part les règles de suppression des consonnes par le DAP, d'autre part la règle phonologique du /ə/. Prenons comme exemple le mot : *vêtement*/vɛtəmɑ̃/. La règle phonologique va le dériver en : CO/VO/CO/VO/CO/VO et CO/VO/CO/CO/VO. En appliquant les règles du DAP – reconnaissance optionnelle des consonnes non inter-vocaliques, dans ce cas le /v/, et présence obligatoire d'uniquement la consonne la plus fermée d'un groupe consonantique identifiable, dans ce cas le /t/ du groupe /tm/ – nous obtenons les images suivantes :

CO/VO/CO/VO/CO/VO,
CO/VO/CO/CO/VO,
CO/VO/CO/VO,
VO/CO/VO/CO/VO,
VO/CO/CO/VO,
VO/CO/VO.

Ces images représentent les formes sous lesquelles peut apparaître le mot /vɛtəmɑ̃ / dans l'ensemble des hypothèses déduites d'un treillis le représentant.
Le lien est maintenant établi entre les hypothèses déduites du treillis phonétique et les items du lexique. Pour permettre un accès optimisé aux informations du dictionnaire, celui-ci est codé dans une table dont les clefs d'accès sont les différentes images déduites des décompositions phonétiques.

4.3. GRAPHE PHONÉTIQUE

Afin de lier les images calculées lors de la génération d'hypothèses sur la structure du mot prononcé et les informations phonétiques produites par le DAP, nous allons réunir l'ensemble de ces données sous la forme d'un graphe acyclique. Les états de ce graphe sont les familles de phonèmes préalablement segmentées. Les différents chemins reliant ces états représentent toutes les structures possibles du mot prononcé selon les images VO/CO calculées. Tous les traitements lexicaux utiliseront ce graphe plutôt que le treillis phonétique. Les phénomènes d'insertions et de suppressions inhérents au DAP sont ainsi pris en compte, non plus au niveau de chaque mot, mais lors de la phase d'analyse du treillis en déterminant les chemins possibles dans le graphe phonétique. L'accès lexical se trouve ainsi optimisé par la prise en compte globale de ces phénomènes (figure 1).

4.4. LOCALISATION

Une fois le codage en graphe réalisé, se pose le problème de la méthode d'accès. Le manque d'information sur les données traitées et le nombre important de possibilités issues des modules en amont empêchent une approche ascendante du problème. De plus, la faible taille des vocabulaires utilisés dans les SDMC (systèmes de détection de mots cibles) (quelques dizaines de mots au maximum) rend raisonnable une stratégie purement descendante utilisant les informations codées dans le dictionnaire pour vérifier leur présence au niveau des données précédemment calculées.

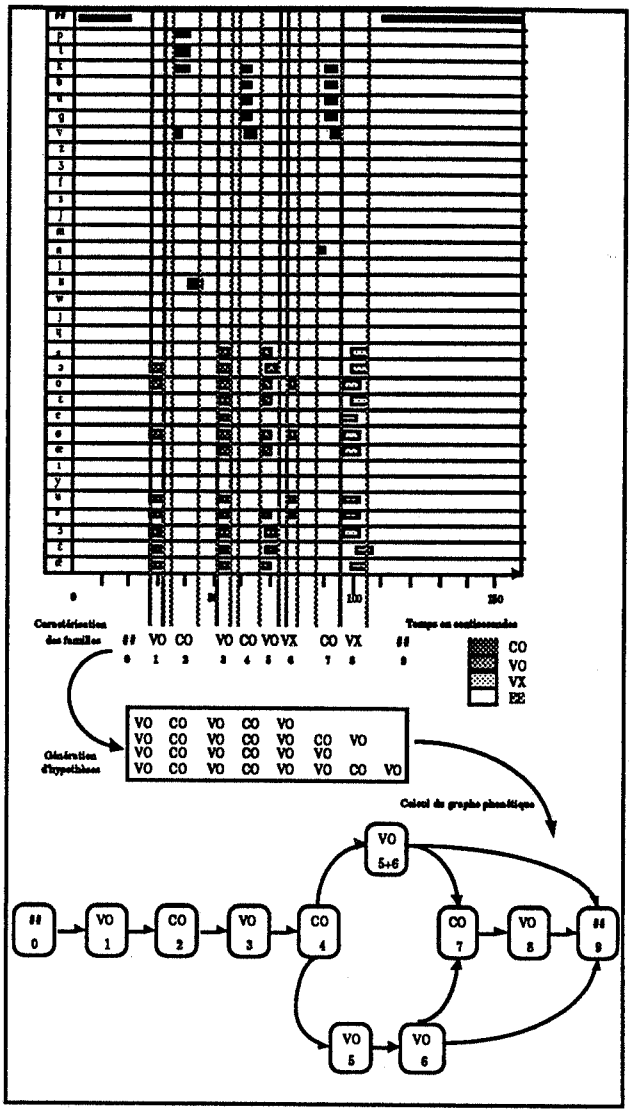


Figure 1 : exemple de traitement du mot *contrebande* avec segmentation du treillis en familles de phonèmes, caractérisation de ces familles et calcul du graphe phonétique regroupant toutes ces informations.

Le premier traitement dans la recherche des mots cibles consiste à parcourir le graphe avec la liste des images de la table d'index du lexique. Pour chaque image de la table, on cherche tous les sous-graphes possibles qui peuvent la représenter. L'algorithme consiste à utiliser la table de transition du graphe en avançant suivant les chaînes VO/CO des images traitées.
Nous disposons ainsi, pour chaque mot-clé compatible avec le treillis, d'une ou plusieurs zones de signal segmentées en familles de phonèmes. Ce premier traitement permet de faire un tri parmi les informations disponibles afin de distinguer les zones considérées comme du "bruit" et les zones susceptibles de contenir un des mots recherchés.

4.5. ÉVALUATION DES HYPOTHÈSES ÉMISES

Une fois la phase de segmentation et de localisation terminée, les hypothèses émises vont être examinées afin de vérifier la présence des phonèmes obligatoires, au sens du DAP, dans les familles correspondant aux décompositions phonétiques des items sélectionnés. Les items restant sont alors évalués après avoir cherché les éventuels phonèmes manquant dans le treillis. Cette évaluation correspond à la moyenne des notes des unités phonétiques composant le mot examiné.

Le résultat est donné sous la forme d'une cohorte de mots-clés probables. Chaque proposition est valuée mais cette note ne permet pas de classer les hypothèses entre elles. En effet les mots proposés, contrairement à la reconnaissance de mots isolés, ne couvrent pas tous la même portion de signal. Certains se chevauchent et d'autres sont disjoints. De plus nous ne savons pas combien de mots du lexique de mots-clés ont été prononcés, ni même s'il y en a. La notation a donc ici pour but d'éliminer les hypothèses trop improbables, c'est à dire dont la note est supérieure à un seuil. Ce seuil est calculé dynamiquement car la méthode de valuation d'un mot utilise des distances spectrales qui ne sont pas normalisées. Nous ne pouvons donc définir à priori un seuil de confiance puisque la note de la réalisation d'un mot est liée aux conditions d'enregistrement et à la qualité des références du locuteur. Nous avons donc été amenés à déterminer le seuil de manière dynamique en considérant que si le mot proposé a une note trop éloignée des meilleures propositions phonétiques de la zone où il se trouve, il doit être rejeté. Pour cela, nous calculons la note moyenne des meilleures unités phonétiques reconnues par le DAP qui sont présentes dans les familles phonétiques décrivant l'item proposé. Ce seuil de confiance permet ainsi de mettre en rapport la qualité de la reconnaissance du mot avec celle des unités phonétiques reconnues par le DAP sur la zone de signal correspondante.

5. STRATÉGIE DE REJET

Comme nous l'avons dit précédemment, l'une des principales difficultés des SDMC est de déterminer une stratégie de rejet des hypothèses erronées. En effet dans le cadre d'un dialogue il est généralement plus efficace de ne pas considérer des mots-clés mal reconnus – conduisant l'utilisateur à reformuler sa requête – plutôt que de diriger le système vers des choix non voulus, toujours plus coûteux à corriger et plus destabilisants pour le locuteur. Nous avons vu au paragraphe précédent que l'évaluation des hypothèses par les distances spectrales permettait de supprimer quelques-unes d'entre elles. Cependant ce filtrage reste insuffisant et nous avons été amenés à employer d'autres méthodes afin de valider les hypothèses émises. Pour cela nous utilisons des règles phonétiques contextuelles qui ont été mises au point dans le cadre du DAP descendant.

5.1. PRÉSENTATION DES TRAITS RETENUS

Les traits acoustico-phonétiques que nous utilisons sont inspirés de ceux proposés par Jakobson pour la classification des phonèmes par opposition mutuelle fondé sur une corrélation plus ou moins directe de l'acoustique d'un phonème et de ses propriétés articulatoires [Jakobson, 63; Méloni, 82; Gilles, 93].

Ainsi, dans notre système l'évaluation d'un phonème revient à valider ses propres traits dans le contexte phonémique où il est énoncé, à l'aide d'un ensemble de règles binaires associées à chaque trait. Afin de valider ou d'infirmer les règles définissant les traits, nous avons été amenés à les vérifier sur un important corpus de test. Nous disposons ainsi d'un seuil de confiance pour chacune d'entre elles. Ces seuils nous permettent de définir diverses stratégies tenant compte des incertitudes de la vérification.

5.2. STRATÉGIE DE VÉRIFICATION

Une fois les traits définis, il nous reste à déterminer une stratégie intégrant chacune des règles. Son rôle est de valider les items de la cohorte finale. Elle est basée sur les deux règles suivantes :

- un phonème est validé si tous les traits retenus pour son évaluation ont été confirmés par les règles descendantes,
- un mot est supprimé de la cohorte de résultats si le pourcentage de ses phonèmes rejetés par rapport au nombre de phonèmes le constituant est supérieur à un seuil S.

Ce pourcentage S permet de faire varier la "sévérité" du filtrage descendant en fonction de l'application visée. Nous avons testé diverses valeurs de cette constante. La spécificité de la reconnaissance par mots-clés justifie le risque de rejet important d'items mal reconnus.

6. PREMIÈRES ÉVALUATIONS

Un système de test, baptisé *SPOT*, a été réalisé afin de valider les stratégies de rejet d'hypothèses présentées précédemment. Pour évaluer correctement un SDMC, il faudrait donc disposer d'un très grand corpus (plusieurs heures) constitué, par exemple, de requêtes d'utilisateurs concernant la consultation d'une base de données énoncées en parole spontanée. Ne disposant pas d'un tel corpus pour le français (des corpus de ce type existent pour la langue anglaise, par exemple Switchboard [Godfrey, 92]), nous avons décidé, dans un premier temps, d'utiliser un corpus qui a été mis au point au LIA et qui concerne le traitement des mots isolés. Nos tests ont été effectués en prenant comme lexique de mots-clés 20 items de ce corpus. Ces items ont été choisis au hasard avec une contrainte sur leur longueur : ils ont, en moyenne, 6 phonèmes afin de pouvoir être reconnus sur la plus grande partie des données enregistrées. Le corpus de tests se compose de la réalisation de 3600 mots par trois locuteurs, soit environ 2h30 de signal.

Ces tests sont évidemment partiels, car il ne permettent pas suffisamment de juger le comportement de notre algorithme de recherche sur de la parole continue et

spontanée. Néanmoins, ils ont le mérite d'évaluer le rôle de notre processus de rejet en mettant en rapport le nombre de fausses alertes et le taux de reconnaissance des items recherchés.
Les résultats sont présentés (table 1) selon trois phases :

- Phase 1 : résultats après l'étape d'accès lexical composée de la localisation et de la valuation des hypothèses émises.
- Phase 2 : résultats après l'utilisation du seuil de rejet dynamique calculé sur les familles des hypothèses restantes.
- Phase 3 : résultats donnés après le filtrage contextuel descendant. Nous donnerons les résultats pour deux valeurs (50% et 40%) de la constante de rejet S.

Notons qu'un mot est considéré comme reconnu uniquement s'il est classé premier dans la cohorte résultat.

	% de mots reconnus	nombre de fausses alertes par heure de parole et par item
Phase 1	100%	175
Phase 2	96%	38
Phase 3		
S = 50%	86%	20
S = 40%	68%	11

Table 1 : résultats des tests selon le type de filtrage

Un corpus est en cours de développement afin de tester nos méthodes sur de la parole continue. Les premiers tests effectués sur 40 phrases prononcées par deux locuteurs ont permis de valider les choix effectués.

7. CONCLUSION

Ces résultats mettent en valeur d'une part la nécessité d'un processus de filtrage, d'autre part le coût important en taux de reconnaissance de la réduction du nombre de fausses alertes. Le module de filtrage descendant présente ici des résultats intéressants. Le fait de pouvoir jouer sur le type des traits utilisés ainsi que sur leur prise en compte au niveau du mot permet d'élaborer de nombreuses stratégies en adaptant le taux de reconnaissance requis au nombre de fausses alertes acceptables.
Le système SPOT doit être testé maintenant sur un corpus plus représentatif des conditions d'utilisation d'un module de word spotting. Les perspectives de ce travail sont donc la mise au point d'un environnement de test appliqué à un domaine particulier et comprenant des processus de gestion du dialogue afin d'évaluer de manière globale les performances d'un tel système.

BIBLIOGRAPHIE :

[Béchet, 94] F. Béchet, Un système de traitement de connaissances phonétiques et lexicales : application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu, Thèse de l'Université d'Avignon et des Pays de Vaucluse.

[Boite, 93] J.M. Boite, H. Bourlard, B. D'Hoore, M. Haesen : A new approach towards keyword spotting, 3rd Eurospeech, Berlin, pp. 1273-76, 1993.

[Gilles, 93] P. Gilles ; Décodage phonétique de la parole et adaptation au locuteur, thèse de l'université d'Avignon et des pays de Vaucluse, 1993.

[Godfrey, 92] J. Godfrey, E. Holliman, J. McDaniel, SWITCHBOARD : Telephone Speech Corpus for Research Development, Proc. IEEE ICASP, pp. I-517-520, 1992.

[Jakobson, 63] R. Jakobson ; Essai de linguistique générale, Editions de minuit, 1963.

[Lleida, 93] E. Lleida, J.B, Marino, J. Salavedra, A. Bonafonte, E. Monte, A. Martinez ; Out-of-vocabulary word modelling and rejection for keyword spotting, 3rd Eurospeech, Berlin, pp. 1265-68, 1993.

[Méloni, 82] H. Méloni ; Etude et réalisation d'un système de reconnaissance automatique de la parole continue ; thèse de doctorat d'état, Aix-Marseille II, 1982.

[Méloni, 91] H. Méloni, P. Gilles ; Décodage acoustico-phonétique ascendant, Traitement du Signal, Vol. 8, No 2, pp.107-114, 1991.

[Rohlicek, 93] J.R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, M. Siu ; Phonetic training and language modeling for word spotting, Proc. Int. Conf. on Acoust., Speech and Sig. Processing, April 1993.

[Rose, 90] R.C. Rose, D.B. Paul ; A Hidden Markov Model based keyword recognition system, Proc. ICASSP, April 90.

[Rose, 93] R.C. Rose ; Definition of subword acoustic units for word spotting, 3rd Eurospeech, Berlin, pp. 1049-52, 1993.

[Weintraub, 93] M. Weintraub ; Keyword spotting using SRI's decipher large vocabulary speech recognition system, Proc. Int. Conf. on Acoust., Speech and Sig. Processing, April 1993.

[Wilpon, 90] J.G. Wilpon, L.R. Rabiner, C.H. Lee, E.R. Goldman, Automatic recognition of keywords in unconstrained speech using hiden Markov models, IEEE Trans. on Acoust. Speech and Signal Proc., 38(11):1870-1878, 1990.

INVERSION DU MODÈLE DRM POUR LA RECONNAISSANCE DE LA PAROLE : APPLICATION À L'IDENTIFICATION DE DIPHONES VOCALIQUES

LAURENCE CANDILLE *, HENRI MÉLONI *, THIERRY SPRIET *, RENÉ CARRÉ **

* LABORATOIRE D'INFORMATIQUE, FACULTÉ DES SCIENCES, AVIGNON

** ENST - DPT SYSTÈME PARIS

Résumé

Nous présentons une utilisation du Modèle à Régions Distinctives (DRM) pour la reconnaissance de diphones vocaliques. La méthode proposée prend en compte des informations relatives au locuteur, au contexte et à la dynamique des paramètres dans la transition V1-V2. Les résultats obtenus sont encourageants et justifient que les travaux soient poursuivis pour identifier les sons au moyen de modèles de production.

Mots clés: Inversion de modèles de production, reconnaissance de la parole

1 - INTRODUCTION

La capacité réduite des modèles de production de parole à rendre précisément tous les processus de la communication vocale impose une limitation de leur utilisation dans le cadre de la reconnaissance [Méloni, 1994]. Une première restriction consiste à ne prendre en compte que les informations acoustiques, en négligeant les connaissances que l'on pourrait extraire d'autres modes d'expression des phénomènes articulatoires (mouvements des lèvres, évolution effective des articulateurs, variations des influx de commande des organes de la phonation, etc.).

Au cours d'un dialogue, la compréhension d'un message vocal s'effectue sur plusieurs plans et prend en compte le processus complet de la production (décodage phonétique, interprétation des variations de la source, évaluation des altérations psycho-physiques, etc.). Même en limitant les traitements au seul décodage phonétique, les modèles que nous envisageons d'utiliser [Carré, 1992; Maeda, 1979] ne permettent pas de considérer toutes les informations utiles pour l'identification des sons (sources de production, modes et lieux articulatoires, stratégies de commande des articulateurs, etc.). Afin de vérifier l'opportunité d'employer ces techniques pour la reconnaissance automatique de la parole, nous avons réalisé des expériences avec un modèle simple dont les commandes peuvent être facilement manipulées (modèle DRM [Carré, 1992]).

Les premiers travaux que nous présentons ont abouti à l'identification de séquences de voyelles orales par comparaison des transitions de formants d'un locuteur avec celles proposées par le modèle DRM. Les résultats obtenus font apparaître que, si l'on choisit convenablement les configurations du modèle pour les voyelles, les informations disponibles dans la phase transitoire entre deux phonèmes améliorent significativement la reconnaissance par rapport à une évaluation fondée exclusivement sur les données disponibles pour les noyaux vocaliques. Les configurations optimales sont obtenues lorsque l'on adapte les représentants en fonction du locuteur et du contexte phonétique.

2 - LE MODÈLE DRM

Le modèle constitué de régions distinctives (DRM) est un modèle de fonctions d'aire. Les régions ont des longueurs fixes (égales à un pourcentage de la longueur totale du conduit) et donc les commande du modèle sont essentiellement transversales. Ce modèle réalise une transformation fonction d'aire / acoustique particulièrement intéressante :

- les plages de variation des formants sont les plus grandes possibles,
- les trajectoires formantiques dans le plan F1/F2 sont monotones dans le cas de transitions V1-V2,
- la relation est pseudo-orthogonale,
- les commandes sont efficaces (critère de minimum d'énergie) c.-à-d. qu'une petite déformation de la fonction d'aire correspond à une grande variation acoustique,
- les commandes de déformaton de la fonction d'aire sont simples (rectilignes) et en nombre réduit,
- par superposition d'un geste consonantique sur une transition V1-V2, on a pu obtenir des productions V1CV2 tout à fait correctes avec un nombre réduit de commandes [Carré, Chennouck, Mrayati, 1991] : à partir de la configuration V1, deux ensembles de commandes sont réalisés, l'un pour aller vers la voyelle V2 (avec un maximum de deux commandes – sur la constriction et sur l'ouverture aux lèvres), l'autre assurant la réalisation de la consonne (par occlusion sur une région). On a donc une co-production syllabique. Le modèle DRM, avec une telle commande, est un bon modèle de coarticulation.

En reconnaissance, la simplicité du modèle et le nombre réduit de ses paramètres, devrait permettre d'élaborer une

stratégie simple pour une procédure d'inversion. En revanche, à cause de sa simplicité et de son caractère normatif, le modèle ne prétend pas à pouvoir représenter toutes les situations (variations inter et intralocuteurs par exemple). Selon la philosophie du modèle, la reconnaissance est un processus essentiellement dynamique : reconnaissance de "gestes" de transitions vers une voyelle, reconnaissance de gestes d'occlusion. Pour ceci, il nest pas nécessaire de mettre en œuvre une inversion mais de détecter directement ces gestes, ces tâches élémentaires de modification de la fonction d'aire.

3 - INVERSION DU MODÈLE DRM

Notre travail ne vise pas à valider le modèle DRM et, dans un premier temps, afin d'évaluer ses possibilités d'inversion pour la reconnaissance de la parole, nous avons appliqué ce processus à l'identification, pour plusieurs locuteurs, de diphones constitués de voyelles orales. L'évaluation de l'adéquation entre le signal vocal et les paramètres produits par le modèle nécessite une paramétrisation appropriée, la définition d'une mesure précise du degré de coïncidence, un processus d'adaptation au locuteur et une stratégie efficace de reconnaissance qui prenne en compte les effets de la coarticulation. De plus, afin de réduire l'ensemble des configurations que l'on peut associer à un vecteur de paramètres acoustiques, nous avons imposé des contraintes qui limitent et lient les variations des sections des régions du DRM.

L'inversion de modèles articulatoires ne peut vraisemblablement pas apporter d'amélioration significative pour l'interprétation statique de paramètres acoustiques, mais cette technique peut s'avérer particulièrement efficace dans l'étude des variations temporelles du signal vocal. Nous avons donc choisi d'utiliser toutes les informations disponibles dans les transitions entre les deux voyelles pour identifier les phonèmes.

3.1 - Contraintes imposées au modèle

Soient $R_{i(i=1..8)}$ les huits régions du modèle (R_8 correspond à l'aire aux lèvres). Nous avons imposé les contraintes suivantes:

$R_1=2\ cm^2$, $R_2=1/2.(R_1+R_3)$, $R_7=1/2.(R_6+R_8)$.

R_i varient de 0,5 à 11 cm^2

Par ailleurs il y a deux possibilités de variation des régions:

- R_3 et R_4 varient simultanément dans le même sens, R_5 et R_6 aussi. De plus ces deux couples de régions varient simultanément, en sens contraire et dans les mêmes proportions.

- dans le deuxième cas c'est parail pour les couples R_3 et R_6, et, R_4 et R_5.

3.2 - Paramètres utilisés

Afin de faciliter la mesure d'une distance entre les paramètres extraits du signal et ceux proposés par le modèle, nous n'avons conservé que les valeurs fréquentielles des 3 premiers formants. Si l'on peut attendre un gain d'information par l'utilisation des bandes passantes et des énergies associées aux formants, l'introduction de ces paramètres dans le processus de comparaison implique une pondération perceptive des différents attributs.

Les valeurs des formants sont calculées automatiquement dans le signal de parole au moyen d'un lissage spectral effectué à partir de coefficients de prédiction linéaire. Le nombre de coefficients est choisi en fonction des voyelles de manière à faire apparaître les pôles rapprochés (F1 et F2 pour [u], F2 et F3 pour [y], F3 et F4 pour [i]). Toutefois, dans certains contextes, il est délicat d'extraire précisément les formants à partir du signal vocal. On peut envisager, pour pallier ces difficultés d'employer d'autres types de paramètres tels qu'un modèle simplifié de représentation spectrale à deux formants [Carlson, 1975] ou des indices acoustiques représentatifs de rapports d'énergie entre canaux spectraux répartis suivant une échelle qui prend en compte les phénomènes de perception [Caelen, 1989].

Si, dans la plupart des cas, le calcul des valeurs des formants ne pose pas de problème majeur sur les zones stables des voyelles, il peut être plus délicat d'évaluer ces paramètres dans certains types de transitions (exemple: le croisement de formants dans la séquence [u i] de la figure 1). Il pourrait être envisagé pour résoudre ces difficultés une normalisation des données formantiques par affiliation directe des résonances aux cavités [Fant, 1975; Payan, 1992].

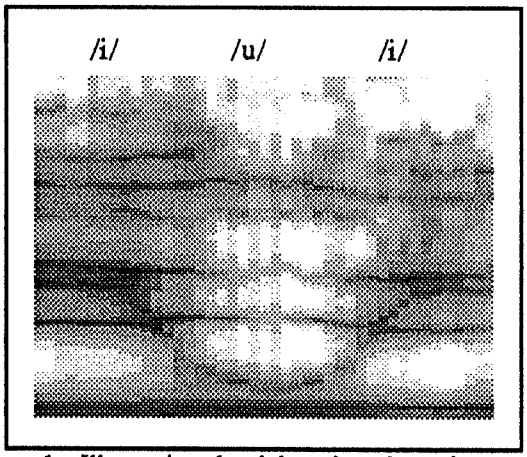


Figure 1 : Illustration du phénomène de croisement de formants dans la séquence [i u i].

3.3 - Mesure de l'erreur de modélisation

Pour évaluer la correspondance entre des hypothèses phonétiques proposées par le modèle et les paramètres mesurés sur le signal, nous avons choisi de comparer les valeurs des formants sur la partie transitoire de la séquence V1-V2. Pour chaque formant, et pour chaque trame de la transition, la distance entre la valeur proposée par le modèle et celle qui lui est associée dans le signal est égale au nombre de bandes critiques qui les séparent. Ce calcul permet de donner une importance perceptuellement équivalente aux trois formants. La

distance totale sur la transition correspond à la somme des distances sur chaque formant de toutes les trames.

3.4 - Adaptation au locuteur

L'espace des variations des formants produits par le modèle doit contenir l'espace des variations des formants calculés pour un locuteur. Cette contrainte n'étant généralement pas vérifiée, il convient soit de modifier certaines caractéristiques du modèle, soit de normaliser les paramètres du locuteur. Les particularités du modèle DRM facilitent son adaptation au locuteur par un choix de la longueur totale du tube et par un décalage de l'axe de symétrie.

Une variation de la longueur du tube induit une variation inverse des valeurs des formants. Le modèle DRM est à l'origine symétrique dans sa structure pour qu'une petite déformation de la fonction d'aire corresponde à une grande variation acoustique. Si les voyelles sont décrites pour un modèle prenant en compte les deux premiers formants la plage du second formant est maximale au détriment de la plage du troisième formant. Un décalage limité de l'axe de symétrie va entraîner une diminution de la plage du second formant et en revanche une augmentation de la plage du troisième formant. Cette situation permet de décrire en particulier la distinction entre [i] et [y] qui pour beaucoup de locuteurs est caractérisée uniquement par la valeur de F3. On pourrait dire que par rapport à la norme ce déplacement de l'axe de symétrie peut être une caractéristique du locuteur. Cette adaptation ne remet pas en cause les stratégies et le nombre de commandes du modèle (ils seraient remis en cause si l'on changeait l'axe de symétrie pour chaque voyelle).

3.5 - Stratégie de reconnaissance

Les configurations standard du modèle DRM proposées pour représenter chaque voyelle correspondent à des valeurs de formants parfois assez éloignées de celle que l'on mesure pour un locuteur donné (figure 2). De plus, les formants d'une même voyelle d'un locuteur varient significativement en fonction du contexte. Nous avons donc choisi de mémoriser, pour la reconnaissance des diphones vocaliques, des configurations qui dépendent du locuteur et du contexte.

Deux types de stratégies de commandes permettent de passer d'une configuration à l'autre par déplacement transversal ou longitudinal de la constriction. La commande longitudinale n'est utilisée que dans le cas du passage d'une voyelle centrale avec les lèvres fermées à une voyelle dont la constriction est arrière ou avant avec les lèvres ouvertes (il faut considérer ces caractéristiques des phonèmes par rapport au modèle).

Le passage d'une voyelle à une autre peut se faire par interpolation linéaire, logarithmique ou quadratique de la valeur des sections. Les différences constatées entre ses trois stratégies ne justifient pas de les distinguer pour la reconnaissance.

Nous envisageons de désynchroniser l'interpolation aux lèvres par rapport aux commandes sur la constriction, mais la prise en compte de ce phénomène ajoute une variable au système qui alourdit considérablement les traitements.

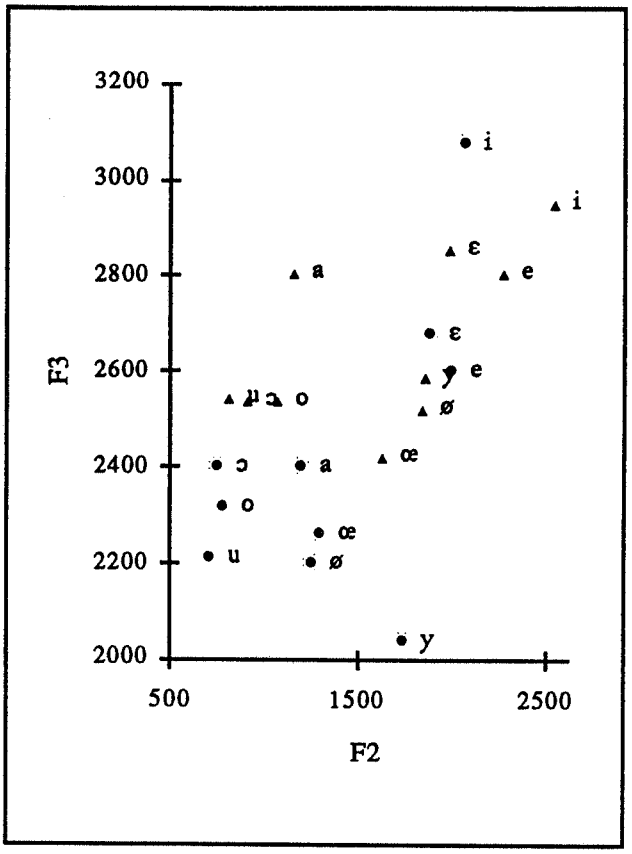


Figure 2 : Comparaison des valeurs formantiques des voyelles standard du modèle et des valeurs formantiques calculées pour un locuteur, dans le plan F2-F3. Les voyelles standards sont représentées par des triangles, celles du locuteur par des points.

3.5.1 - Apprentissage des modèles

L'adaptation de la longueur du tube au locuteur s'effectue à partir d'un ajustement des valeurs des formants pour la position neutre du modèle DRM (correspondant à la voyelle [œ]).

Pour le déplacement de l'axe de symétrie, nous produisons – dans les limites des variations du modèle – un ensemble important de couples associant une configuration donnée aux formants correspondants. À partir des valeurs moyennes des formants pour les dix voyelles orales d'un locuteur, et à l'aide de la mesure définie précédemment, nous obtenons dix configurations de référence. Nous choisissons la valeur du déplacement de l'axe de symétrie correspondant à l'ensemble des couples qui ont permis d'obtenir les configurations dont les paramètres acoustiques minimisent les distances à l'ensemble des phonèmes du locuteur. Dans le processus de reconnaissance la longueur du tube et la position de l'axe de symétrie sont donc fixées en fonction du locuteur.

Pour un locuteur donné et une longueur du tube fixée à 18,5 cm, la figure 3 illustre l'amélioration des résultats acoustiques dans le plan F2-F3 lorsque l'on déplace l'axe de symétrie de 1 cm vers les lèvres par rapport aux résultats produits par le tube normal. La figure 4 permet de comparer, pour les trois voyelles cardinales ([i], [a], [u]), les configurations du modèle DRM et des fonctions d'aires proposées par Fant (Fant 1960).

Afin de prendre en compte les effets de la coarticulation, nous envisageons d'utiliser plusieurs configurations de référence pour une voyelle.

3.5.2 - Reconnaissance des séquences V1-V2

S'agissant d'une étape de validation de l'inversion d'un modèle de production pour la reconnaissance de la parole, nous avons soumis au processus d'identification des séquences de vecteurs de formants correspondant à des réalisations de diphones V1-V2 dont les limites temporelles et les valeurs des paramètres ont été obtenues semi-automatiquement (il n'y a donc pas d'erreur de segmentation ni d'affectation des formants). Dans le cas particulier de séquences V1-V2 isolées la procédure est très simplifiée et pratiquement automatique.

Pour chacune des deux voyelles du diphone, nous choisissons les trois configurations de référence les plus vraisemblables. Plusieurs trajectoires sont produites par le modèle entre ces différentes configurations et au moyen des diverses commandes disponibles. La distance à une séquence "naturelle" de vecteurs de formants est comparée sur toute la transition aux trajectoires proposées par le modèle, le résultat de l'identification est constitué des références qui ont permis de suivre au mieux les paramètres extraits du signal. Cette approche permet une interprétation phonétique directe de la séquence à décoder. Il s'agit donc d'une reconnaissance effectuée sous les contraintes suivantes :

- modèles adaptés au locuteur (longueur du tube, symétrie),
- prise en compte d'informations dynamiques sur la transition V1-V2.

4 - RÉSULTATS

Les premiers essais de recherche statique de configurations du modèle DRM associées aux unités phonétiques ne donnent pas de résultats satisfaisants pour une utilisation en reconnaissance [Jospa, 1990; Candille, 1993]; il est donc nécessaire de prendre en compte les informations dynamiques. Par ailleurs, le processus de reconnaissance que nous avons employé donne des résultats également médiocres si on utilise les configurations standard des voyelles; l'adaptation au locuteur est indispensable. De plus, le choix de configurations dépendantes du contexte affine les mesures et améliore significativement les résultats.

Les premiers tests ont été effectués sur des séquences V1-V2, les résultats obtenus (plus de 90 % de reconnaissance du diphone en première position) sont encourageants mais nous devons évaluer notre technique sur un corpus étendu comportant de nombreux locuteurs.

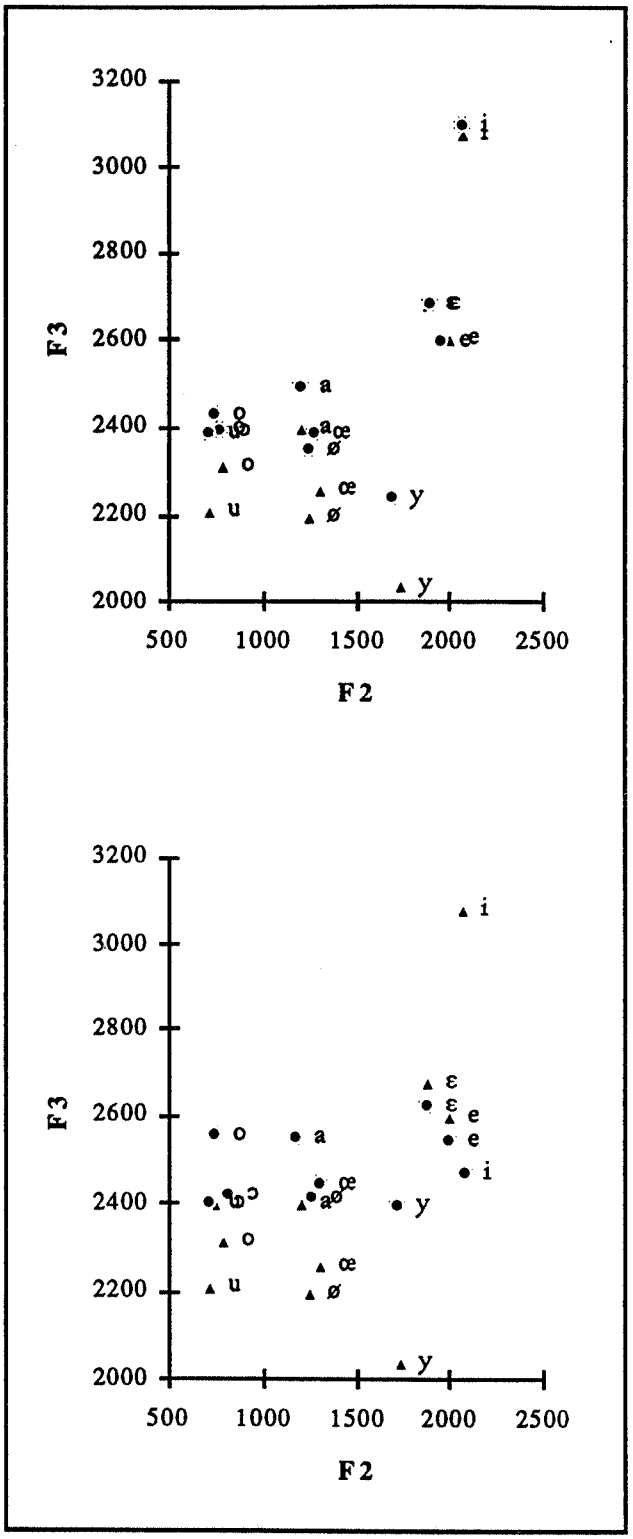


Figure 3 : Comparaisons des résultats acoustiques (dans le plan F2-F3) produits par un tube dont la symétrie à été décalée de 1 cm vers les lèvres (en haut) et par un tube symétrique (en bas). Les phonèmes du locuteur sont représentés par des triangles, ceux du modèle par des points.

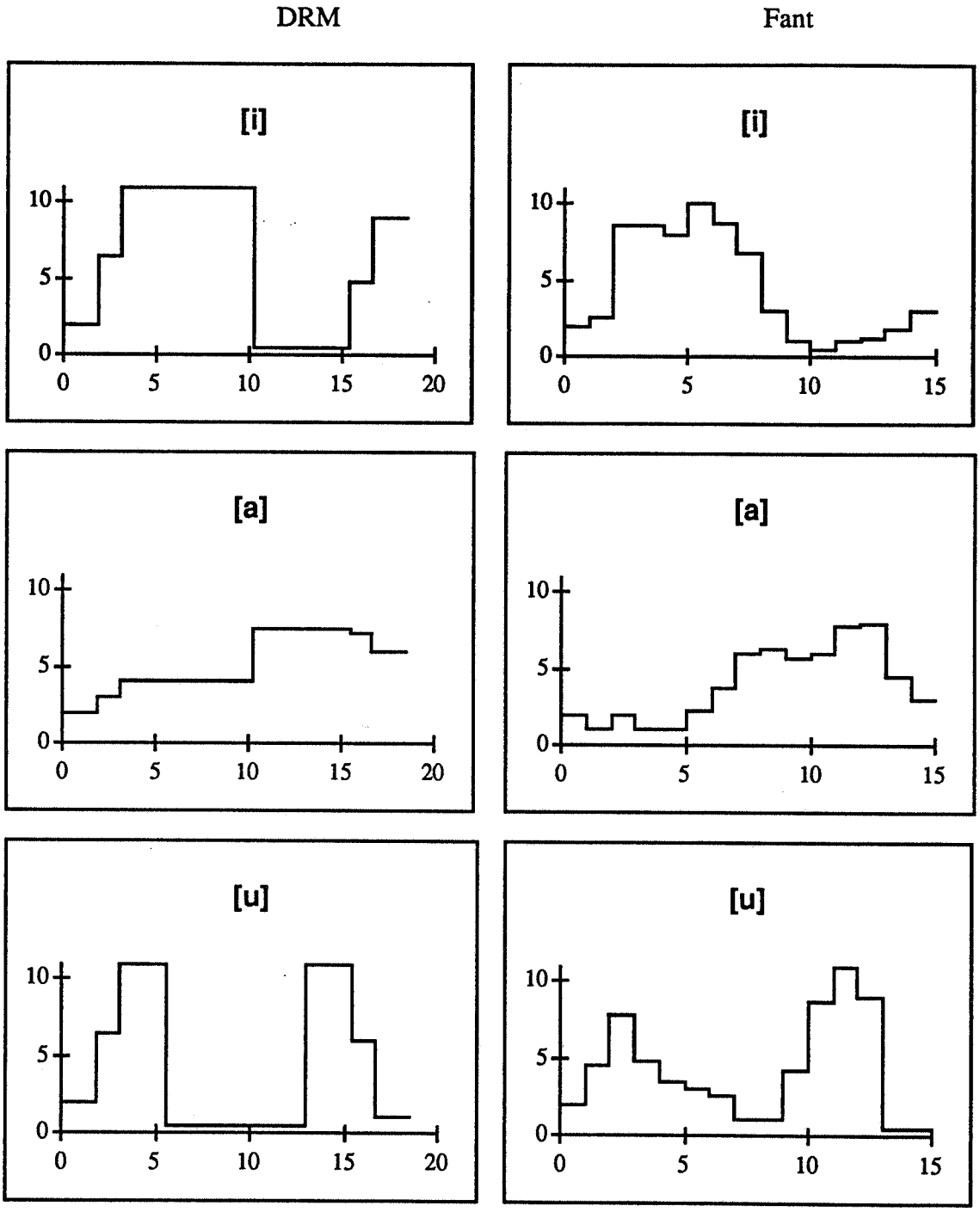


Figure 4 : Comparaison des fonctions d'aires obtenues à l'aide du modèle DRM adapté en longueur et symétrie au locuteur, et celles données par Fant (Fant 1960) pour les voyelles [i], [a] et [u]. L'axe des abscisses représente la distance à la glotte en cm et l'axe des ordonnées l'aire des régions en cm^2.

5 - CONCLUSION

Les premières expériences que nous avons menées à partir de l'inversion du modèle DRM sont encourageantes pour la reconnaissance de la parole au moyen de cette technique. Il convient toutefois d'approfondir les tests et d'appliquer la méthode à l'identification de tous les phonèmes. La reconnaissance des consonnes exige une modélisation de l'ensemble des sources, des modes et des lieux articulatoires. L'état actuel des modèles permet d'envisager essentiellement l'identification du lieu de l'articulation des consonnes; les techniques classiques se révélant relativement efficaces pour la reconnaissance des modes.

Les modèles de la production de parole peuvent être utilisés en reconnaissance comme des réservoirs d'information disponibles pour l'identification des sons et d'autres phénomènes articulatoires. Il serait intéressant d'utiliser de manière optimale d'autres techniques et plusieurs modèles.

6 - RÉFÉRENCES

CAELEN J. (1989), "Indices pour le décodage acoustico-phonétique", In Bothorel A. et al. (Eds), Mélanges de phonétique générale et expérimentale, Strasbourg

CANDILLE L. (1993), "Inversion de modèles articulatoires pour la reconnaissance de la parole", Rapport de DEA, Université d'Avignon et des Pays de Vaucluse

CARLSON R., FANT G., GRANSTROEM B. (1975), "Two-formant models, pitch and vowel perception", In Auditory Analysis and Perception of Speech, Eds G. Fant & M.A.A. Tatham, Academic Press (London)

CARRÉ R., CHENNOUCK, MRAYATI M. (1991), "Vowel-Consonant-vowel transitions : analysis, modeling, and synthesis", *ICSLP 92*, Banff

CARRÉ R., MRAYATI M. (1992), "Distinctive regions in acoustic tubes. Speech production modelling", Journal d'Acoustique, n° 5, pp. 141-159

FANT G. (1960), "Acoustic theory of speech production", Mouton Co.

FANT G. (1975), "Non-uniform vowel normalization", *STL-QPSR*, 2/3 PP. 1-19

MAEDA S. (1979), "An articulatory model of the tongue based on statistical analysis", *JASA*, 65: s22

MÉLONI H., CANDILLE L., SPRIET T., CARRÉ R., PERRIER P., PAYAN Y., LOWENBRUCK H., CAELEN J. (1994), "Les modèles de production et la reconnaissance de la parole : objectifs et difficultés", *Séminaire GDR-PRC CHM Reconnaissance de la parole*, Nancy

MRAYATI M., CARRÉ R. (1992), "The acoustic-area function inversion problem and the distinctive region model", Signal Processing VI : Theories and Applications, J. Vandewalle, R. Boite, M. Moonen, A. Oosterlinck (eds.), Elsevier Sciences Publishers

PAYAN Y. (1992), "Normalisation de l'espace de réalisation acoustique des locuteurs en vue de la synthèse de parole à l'aide d'un robot parlant", Rapport de DEA, INPG, Grenoble

SOQUET A., SAERENS M., JOSPA P. (1991), "Acoustic-articulatory inversion based on a neural controller of a vocal tract model : Further results", *In Proc. of the International Conference on Artificial Neural Networks*, pp. 371-376, Helsinki

TOPOLOGIE DES LOCUTEURS ET PRÉSEGMENTATION POUR LE DÉCODAGE ACOUSTICO-PHONÉTIQUE

C. BARRAS, M.-J. CARATY & C. MONTACIÉ

LAFORIA-IBP, UNIVERSITÉ PARIS 6, CNRS-URA 1095
4 PLACE JUSSIEU - 75252 PARIS CEDEX 5 - FRANCE

Résumé

Le décodage acoustico-phonétiques pose encore des problèmes de segmentation et de variabilité inter-locuteurs. Pour les résoudre, deux méthodes de traitement du signal sont utilisées dans un système fondé sur les modèles actuellement les plus performants: les modèles de Markov cachés. D'une part, les Ruptures de Modèles détectent des discontinuités du signal de parole utilisées pour un contrôle temporel des transitions entre les modèles; d'autre part, les modèles AR-vectoriels permettent la définition d'une topologie des locuteurs qui améliore la qualité de l'ensemble d'apprentissage. Sur la base de données TIMIT, une augmentation significative du taux de reconnaissance est observée lors de l'utilisation conjointe des méthodes. L'apport respectif des méthodes est également analysé.

Mots-clés: décodage acoustico-phonétique, modèles de Markov cachés, ruptures de modèles, modèles AR-vectoriels, segmentation, topologie des locuteurs.

1. INTRODUCTION

Le décodage acoustico-phonétique (DAP) reste une étape fondamentale de la reconnaissance de la parole continue. Les méthodes actuellement les plus performantes dans ce domaine sont fondées sur les modèles de Markov cachés (HMM). Des études sur les espaces de représentation ou sur le choix des modèles permettent une amélioration des performances. Cependant, le DAP est un processus au cours duquel segmentation et identification sont étroitement liées, et les systèmes de reconnaissance à base de HMM ne permettent pas de localiser avec précision les frontières phonétiques. De plus, la variabilité inter-locuteurs rend le choix de l'ensemble d'apprentissage crucial pour une bonne estimation des densités de probabilité des modèles.

Pour résoudre ces problèmes, deux méthodes de traitement du signal sont utilisées: la méthode des Ruptures de Modèles (RM) [And], et les modèles AR-vectoriels. La méthode RM détecte les instants caractéristiques de discontinuités du signal de parole. Ces ruptures permettent un contrôle temporel des transitions entre les modèles phonétiques. Les modèles AR-vectoriels sont utilisés pour définir une topologie des locuteurs dans le but d'améliorer la qualité de l'ensemble d'apprentissage.

A partir d'un système de référence (HMM), un nouveau système de décodage avec transitions contraintes et sélection de l'apprentissage (HMM-TCSA) a été développé. La comparaison de ce nouveau système avec le système de référence a été effectuée sur la base de données TIMIT. Ce nouveau système permet d'augmenter le taux de reconnaissance de plus de 4 %.

2. MÉTHODES D'ANALYSE

Les Ruptures de Modèles sont fondées sur une étude statistique du signal représenté par une suite de zones quasi-stationnaires. Chacune de ces zones est modélisée par un modèle auto-régressif, et les frontières entre ces zones sont les ruptures détectées. Nous proposons l'utilisation de cette information de nature segmentale dans un système à base de HMM. Dans ces systèmes, les modèles de durée sont généralement fondés sur une statistique de durée a priori des phonèmes et ne se réfèrent pas au signal de parole analysé.

2.1. RUPTURES DE MODÈLES

La méthode des Ruptures de Modèles [And] permet de détecter des discontinuités du signal. On considère que le signal $\{s_n\}$ est constitué d'une suite de segments quasi-stationnaires, caractérisés chacun par un modèle auto-régressif d'ordre p:

$$s_n = \sum_{i=1}^{p} a_i s_{n-i} + e_n$$

Les discontinuités sont détectées par la divergence, à l'instant n, entre un modèle à long terme calculé sur une fenêtre croissante, et un modèle à court terme calculé sur une fenêtre glissante. Une statistique sur l'entropie mutuelle des résidus des modèles permet de détecter l'instant de la rupture r_i. La Figure 1 montre un exemple des segments stationnaires entre les ruptures, comparés à l'étiquetage d'origine pour une phrase de la base de données TIMIT.

Dans nos études précédentes [Bar], nous avons appliqué la méthode RM à la segmentation du signal de

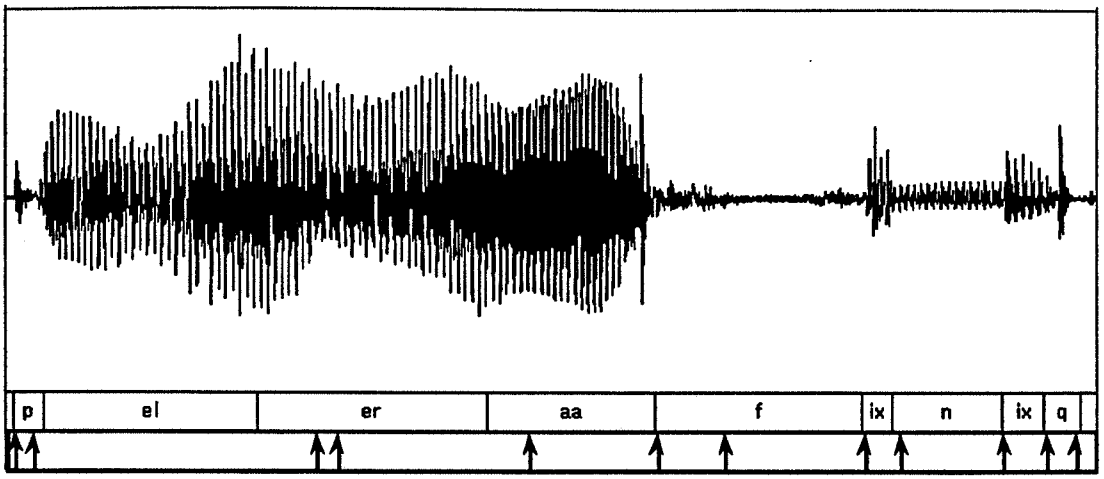


Figure 1. Etiquetage de référence de la phrase: [Elderly peo]ple are often ex[cluded] (première ligne) et positions des ruptures détectées (seconde ligne)

parole. Le nombre de ruptures obtenu s'est avéré être supérieur au nombre de phonèmes à reconnaître, certaines ruptures correspondant à des non-stationnarités internes à un phonème. Cette segmentation engendrée par les ruptures ne peut donc directement être utilisée dans le DAP. Cependant, l'étude statistique de la distribution des ruptures montre une forte corrélation entre la position des ruptures et l'instant de transition des phonèmes. La Figure 2 présente la probabilité d'avoir une transition phonétique conditionnée par la distance de la rupture la plus proche, estimée sur la base de données TIMIT.

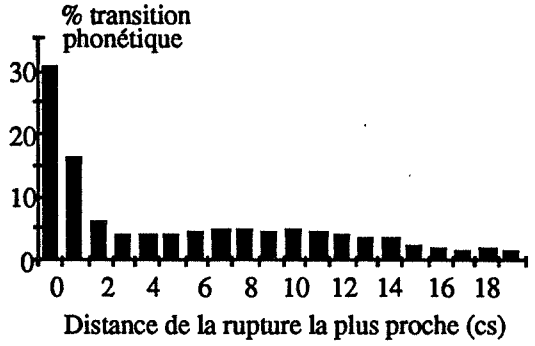


Figure 2: Probabilité de transition phonétique conditionnée à la distance des ruptures

2.2. MODÈLES DE MARKOV CACHÉS À TRANSITIONS CONTRAINTES

Une chaîne de reconnaissance à base de modèles de Markov cachés à densité continue est appliquée à la reconnaissance de la parole continue. Pour chaque étiquette phonétique, un modèle de Markov dépendant des contextes gauche et droit est appris sur les segments correspondants de la base d'apprentissage. Le modèle de Markov choisi pour ces triphones est un modèle de Bakis à trois états. Les probabilités d'émission associées à chaque état sont décrites par une distribution gaussienne. Les probabilités de transitions entre les modèles phonétiques sont estimées sur l'ensemble d'apprentissage (bigramme de modèles).

Les ruptures détectées par la technique RM sont interprétées dans le système de décodage comme des instants de transitions possibles entre les modèles représentant les unités phonétiques. Ainsi, lors de l'étape d'identification utilisant l'algorithme de Viterbi, la probabilité $p(M_j/M_i)$ de transitions entre les modèles M_i et M_j dépend de la distance de la rupture la plus proche (cf. Figure 3). Si cette distance est supérieure à un seuil fixé a priori, la probabilité de transition inter-modèles est multipliée par un coefficient α ($0 \leq \alpha \leq 1$). Plus α est petit, plus la probabilité d'avoir une transition de modèles loin des ruptures est faible (cf. Figure 4). La valeur retenue pour α est celle qui donne les meilleurs résultats sur l'ensemble d'apprentissage.

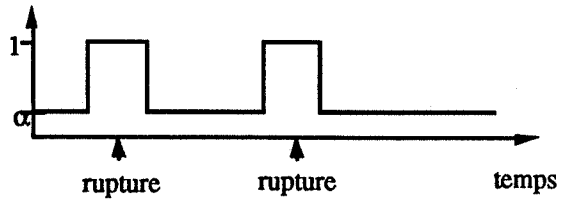


Figure 4. Modification de la probabilité de transition

Le modèle de Markov choisi ici est classique, mais ce protocole peut être appliqué sans aucune difficulté à n'importe quel autre modèle. L'ajout de cette contrainte temporelle doit permettre de diminuer les erreurs.

3. SÉLECTION DE L'ENSEMBLE D'APPRENTISSAGE

La quantité et la qualité des données pose un problème au cours de tout apprentissage. L'augmentation de la quantité des données devrait avoir

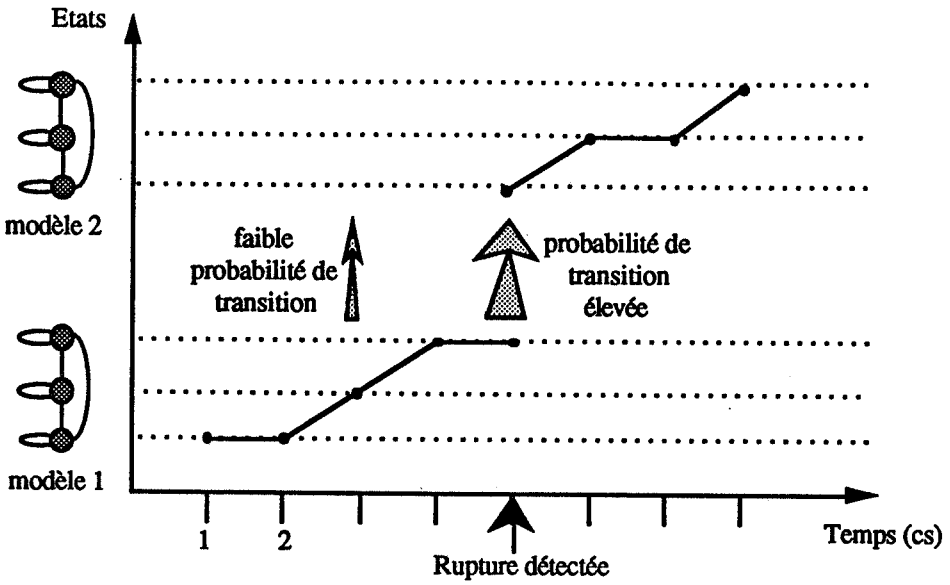


Figure 3. Contrainte des transitions inter-modèles au cours du décodage phonétique

pour résultat une meilleure reconnaissance. Pourtant, un grand nombre de locuteurs entraîne une dispersion des phonèmes dans l'espace acoustique, et rend difficile le choix des frontières entre les classes. L'amélioration de l'ensemble d'apprentissage peut être obtenue par la définition d'une topologie acoustique des locuteurs. Des travaux en vérification et en identification du locuteur [Mon] ont abouti à la définition d'une distance entre locuteurs extrêmement performante, puisqu'elle permet l'identification sans la moindre erreur d'un locuteur parmi les 630 locuteurs de la base de données TIMIT.

3.1. MODÈLES AR-VECTORIELS

Les modèles AR-vectoriels, outil classique d'analyse de signaux complexes, sont utilisés ici pour décrire les vecteurs d'analyse du signal. Soit $\{y_m\}$, (m = 1...N) une succession de N vecteurs spectraux d'ordre p. Leur évolution est décrite par un modèle auto-regressif vectoriel d'ordre q:

$$y_m = \sum_{j=1}^{q} A_j \, y_{m-j} + \varepsilon_m$$

où $\{A_j\}$, (j = 1...q) sont des matrices p x p et $\{\varepsilon_m\}$, (m = 1...N) est un bruit blanc vectoriel de matrice de covariance D.

Les coefficients des matrices A_j sont estimés par l'algorithme de Levinson-Whittle-Robinson. Le critère à minimiser est la trace de la matrice de covariance D. Ces modèles ont été récemment utilisés avec des vecteurs d'analyse de la parole. Appliqués à la reconnaissance du locuteur [Mon], ils sont interprétés en terme de capacités articulatoires du locuteur. La mesure inter-locuteurs choisie IS_I est celle qui donne les meilleurs résultats pour les expériences de reconnaissance du locuteur.

3.2. TOPOLOGIE DES LOCUTEURS

Pour chaque locuteur de l'ensemble de test, la distance IS_I a été utilisée pour choisir un pourcentage des plus proches locuteurs dans l'ensemble d'apprentissage, définissant ainsi une topologie des locuteurs. Pour calculer les modèles AR-vectoriels et la distance IS_I, les paramètres spectraux utilisés sont ceux du décodage acoustico-phonétique (cf. §4.1).

4. EXPÉRIENCES

Les deux systèmes de DAP indépendants du locuteur HMM et HMM-TCSA sont évalués sur la base de données TIMIT. Pour analyser les apports respectifs des deux méthodes de traitement du signal utilisées, des expériences sont aussi menées avec deux variantes du système HM-TCSA: les systèmes HMM-TC et HMM-SA.

4.1. PARAMÈTRES ET MODÈLES

Le signal, échantillonné à 16 kHz sur 16 bits, est représenté par une suite de vecteurs de 12 coefficients MFCC (Mel Frequency Cesptral Coefficients) [Dav]. Ces coefficients sont calculés à partir des énergies extraites de 24 filtres triangulaires répartis sur l'échelle de fréquence Mel. A cette représentation sont ajoutés l'énergie temporelle ainsi que les coefficients différentiels des coefficients MFCC et de l'énergie.

Les systèmes de DAP sont développés à partir de la bibliothèque HTK de l'Université de Cambridge. Pour chaque classe phonétique en contexte gauche et droit, un

modèle de Bakis à 3 états est appris sur les segments de référence de la base de données d'apprentissage. Afin d'estimer les densités sur des données suffisantes, les contextes sont limités à 5 macro-classes phonétiques (i.e. voyelles, consonnes nasales, consonnes liquides, autres consonnes et silences), et les contextes gauche et droit sont appris de manière indépendante. Ainsi, le nombre d'états différents pour tous les modèles de triphones est ramené à environ 500.

4.2. BASE DE DONNÉES ET APPRENTISSAGE

Les expériences ont été effectuées sur la base de données TIMIT. Cette base est une base de référence pour l'évaluation des systèmes de reconnaissance de la parole continue indépendamment du locuteur. Les 630 locuteurs enregistrés (438 hommes et 192 femmes) sont originaires des Etats-Unis et sont classés en huit dialectes de l'anglais américain. Les locuteurs prononcent chacun 10 phrases: 3 phrases différentes pour tous les locuteurs, 5 phrases choisies parmi 450 phrases phonétiquement équilibrées, et 2 phrases, représentatives des variations dialectales, identiques pour tous les locuteurs. Ces deux dernières phrases ne sont pas utilisées pour le DAP. Dans cette base, 462 locuteurs constituent l'ensemble d'apprentissage. Parmi les locuteurs restants, 24 locuteurs (16 hommes et 8 femmes) forment l'ensemble de test retenu. L'étiquetage phonétique d'origine de la base se répartit en 61 classes phonétiques; nous avons retenu une réduction par regroupement d'allophones en 39 classes phonétiques pour l'évaluation de l'identification. Cet étiquetage, proposé par Lee & Hon [Lee] et utilisé ensuite par d'autres chercheurs [Rob] [Nil] [Chi] permet une évaluation comparative des résultats.

La topologie des locuteurs a été obtenue à partir des modèles AR-vectoriels calculés sur les deux phrases dialectales de chaque locuteur; les huit phrases restantes sont disponibles pour le DAP. Pour chaque locuteur de l'ensemble de test, 10% des plus proches locuteurs (i.e., 46 locuteurs) sont utilisés pour apprendre les modèles de ce locuteur.

4.3. RÉSULTATS

Pour chaque phrase de test, les phonèmes identifiés, substitués, omis ou insérés sont calculés à partir d'un alignement dynamique entre la chaîne phonétique de référence et la chaîne phonétique produite par le système de décodage. La somme des taux d'identification, de substitution et d'omission est par définition de 100%. Le taux global de reconnaissance est défini par convention comme la différence du taux d'identification et du taux d'insertion.

Le Tableau 1 présente les résultats comparatifs des systèmes de décodage HMM et HMM-TCSA. On constate une amélioration du taux d'identification ainsi qu'une légère diminution du taux d'insertion, conduisant à une augmentation du taux de reconnaissance (≈ 4 %).

Le tableau 2 présente les résultats comparatifs des systèmes HMM-TC et HMM-SA. Le détail des résultats sur les systèmes HMM-TC et HMM-SA montre que l'amélioration du taux d'identification provient en grande partie de la sélection de l'apprentissage (augmentation de ≈ 4 %), les apports au système des transitions contraintes entre modèles étant principalement dûs à la réduction des insertions (diminution de ≈ 2,5 %), malgré une légère baisse du taux d'identification (diminution de 0,8 %). Globalement, chacun des deux systèmes apporte une amélioration comparable du taux de reconnaissance (augmentation de ≈ 2 %).

Taux (%)	HMM	HMM-TCSA
Identification	65,1 %	68,1 %
Substitution	27,2 %	24,9 %
Omission	7,7 %	7,0 %
Insertion	10,6 %	9,4 %
Reconnaissance	*54,5 %*	*58,7 %*

Tableau 1. Résultats comparatifs des systèmes HMM et HMM-TCSA

Taux (%)	HMM-TC	HMM-SA
Identification	64,3 %	69,2 %
Substitution	26,6 %	24,1 %
Omission	9,1 %	6,1 %
Insertion	8,0 %	12,5 %
Reconnaissance	*56,3 %*	*56,7 %*

Tableau 2. Résultats comparatifs des systèmes HMM-TC et HMM-SA

5. CONCLUSION

Cette étude montre que l'utilisation combinée d'une information segmentale et de la topologie acoustique des locuteurs permet une amélioration significative des résultats. Cependant, si dans ces expériences l'information segmentale a apporté des résultats prometteurs, la prise en compte des Ruptures de Modèles reste encore trop simplifiée. La topologie choisie pour la sélection des locuteurs semble pertinente, et nous encourage à poursuivre dans cette voie; en particulier, l'apprentissage de modèles phonétiques par classe de locuteurs (sans se limiter à l'habituelle classification entre hommes et femmes) semble nécessaire. L'intégration de la topologie des locuteurs et une meilleure utilisation des RM dans un système plus sophistiqué sont donc nos objectifs.

RÉFÉRENCES

[And] R. André-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE Trans. ASSP, vol. 36, pp. 29-40, 1988.

[Bar] C. Barras, M.-J. Caraty, P. Deléglise, C. Montacié, R. André-Obrecht & X. Rodet, "Décomposition Temporelle et Ruptures de Modèles pour le Décodage Acoustico-Phonétique", 19th JEP, pp. 335-340, 1992.

[Dav] Davis & Mermelstein, "Comparaison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences", IEEE ASSP, vol 28, 1980.

[Lee] K.-F. Lee & H.-W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models", IEEE Trans. ASSP, vol. 37, pp. 1641-1648, 1989.

[Rob] T. Robinson, J. Holdsworth, R. Patterson & F. Fallside, "A Comparison of Preprocessors for the Cambridge Recurrent Error Propagation Network Speech Recognition System", ICSLP, pp. 1033-1036, Kobe, 1990.

[Nil] L.T. Niles, "TIMIT Phoneme Recognition using an HMM-derived Recurrent Neural Network", Eurospeech, pp. 559-562, Gènes, 1991.

[Chi] B. Chigier & H. C. Leung, "The Effects of Signal Representations, Phonetic Classification Techniques, and the Telephone Network", ICSLP, pp. 97-100, Banff, 1992.

[Mon] C. Montacié & J.-L. Le Floch, "Discriminant AR-Vector Models for Free-Text Speaker Verification", Eurospeech, pp. 161-164, Berlin, 1993.

REMERCIEMENTS

Nous tenons à remercier R. André-Obrecht pour son logiciel de calcul des Ruptures de Modèles.

RECONNAISSANCE DE MOTS ENCHAINES PAR UNE MÉTHODE HYBRIDE : RÉSEAUX MARKOVIENS ET BASE DE REGLES

Olivier Delemar
Institut de la Communication Parlée, URA CNRS $N°368$
INPG & Université Stendhal
46 Av. Félix Viallet 38031 Grenoble Cedex

8 avril 1994

Résumé

Cet article décrit une méthode hybride de reconnaissance automatique de mots connectés. Elle permet de contraindre le chemin parcouru dans le graphe des états de modèles de Markov cachés en utilisant les informations fournies par un décodage acoustico-phonétique à base de règles. Les premières expériences réalisées par simulation du DAP à base de règles confirment la faisabilité du système et permettent de définir de nouvelles voies dans la détermination de méthodes hybrides.

1 Introduction

L'utilisation conjointe de méthodes globales et analytiques de reconnaissance automatique de la parole a déjà fait l'objet de plusieurs publications dans la littérature, qu'il s'agisse d'une utilisation en parallèle avec confrontation des résultats dans un *tableau noir* (*blackboard*) [6], ou de faire appel aux méthodes globales pour aligner des événements phonétiques prédits par règles [5] [12]. Le modèle hybride décrit dans cet article est basé sur les HMMs *mots connectés* développés à l'Institut de la Communication Parlée dans le cadre du projet *MultiWorks* [3] [11] et le DAP à base de règles du système *DIRA* [10].

Dans le cadre des applications de type *commandes vocales* telles que le système *ICPplan* [1], le choix du mot comme unité lexicale reste bien adapté [7][2]. En effet le vocabulaire généralement restreint n'impose pas le stockage d'un nombre exorbitant de références et la possibilité de poser des contraintes syntaxiques fortes facilite l'enchaînement des modèles par les méthodes classiques de reconnaissance de mots connectés [8]. De plus, les corpus n'étant pas nécessairement phonétiquement équilibrés, le problème du nombre suffisant de répétitions de chaque unité pour l'apprentissage ne se pose pas dans les mêmes termes que pour les modèles phonétiques ou syllabiques.

Cependant, ces modèles restent incapables de détecter la présence de certains macro-traits à caractères dynamiques tels que les occlusives, macro-traits que les systèmes basés connaissances prédisent avec d'assez bonnes performances (de 92% à 95% pour le système *DIRA*)

En se basant sur l'étude des causes d'erreur les plus courantes des modèles HMM et de l'évolution des probabilités des états, cet article propose un système hybride et décrit les résultats obtenus lors de tests préliminaires. Vient ensuite un paragraphe consacré à la comparaison de ce modèle avec d'autres méthodes d'hybridation. Les adaptations et expériences en cours sont alors présentées pour finir par la conclusion.

2 Le système hybride

Les HMMs utilisés sont des modèles linéaires discrets à 5 états dont les transitions autorisées vont d'un état à lui même ou au suivant. La phase d'analyse du signal se fait sur des fenêtres de 20ms glissantes par pas de 10ms sur lesquelles on calcule 15 coefficients cepstraux obtenus par analyse LPC. Une quantification vectorielle selon l'algorithme de Linde, Buzo et Gray classe les vecteurs résultants parmi 128 symboles. L'apprentissage est effectué par l'algorithme de Baum-Welch multi-observations et la reconnaissance par celui de Viterbi [9]. On peut ainsi par *backtracking* remonter le chemin optimal parcouru lors de la reconnaissance d'une occurrence d'un mot donné. De tels chemins sont représentés sur la figure 1.

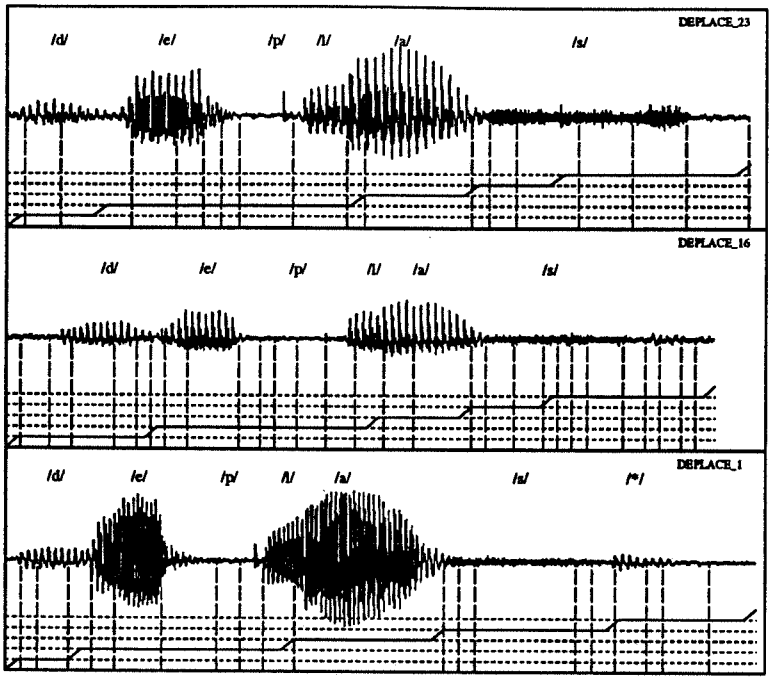


Figure 1: Relation entre graphe des états et segmentation.

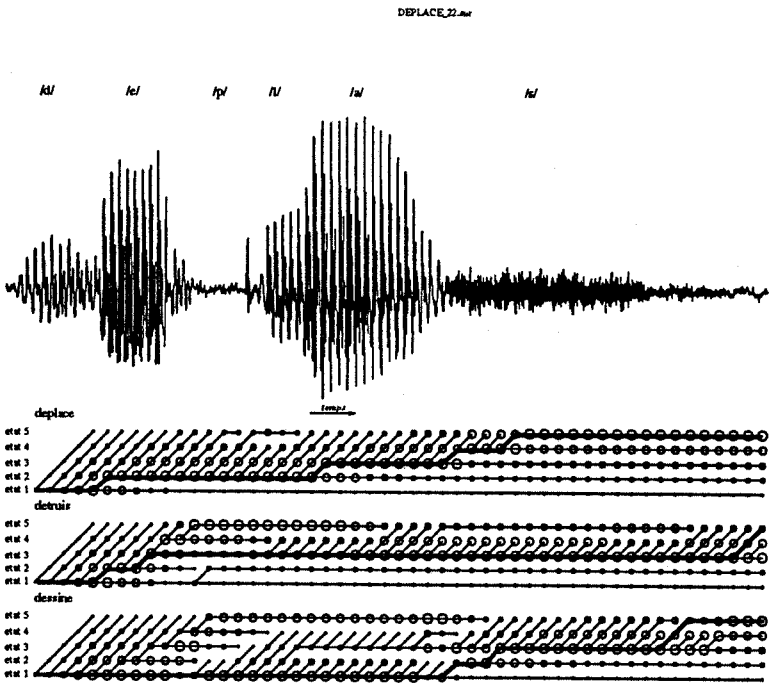


Figure 2: Comportement de chaque état des modèles.

Cette figure représente le chemin parcouru parmi les états lors de l'application de l'algorithme de Viterbi pour trois occurrences du mot *déplace* sur le modèle correspondant. Les changements d'états sont représentés par le passage du trait plein d'une ligne horizontale en pointillés à une autre. Les lignes verticales sont les marques posées par la fonction de présegmentation automatique de *DIRA*. Pour vérifier l'apparente relation que l'on peut observer entre le passage par un état et la présence de phonèmes, une étude sur un corpus étiqueté manuellement de 6 répétitions par 6 locuteurs d'une liste de 80 mots (extrait de *BDSONS*) a montré la présence en moyenne de 1 à 2 phonèmes par état (pour respectivement 49,25% et 44,75% des états).

L'idée première du système hybride est donc de créer à la suite de l'apprentissage une troisième matrice (après les matrices de probabilités d'émission et de transition) représentant le taux de recouvrement de chaque état par chaque macro trait que sait prédire le DAP à règles. Pour cela, chaque occurrence du mot ayant servi à l'apprentissage subit une phase de reconnaissance sur le modèle suivit d'un *backtracking* tandis que l'on procède à une segmentation et à un étiquetage en macro-traits par règles.

Le nombre de trames d'analyse pour lesquelles on a simultanément le passage par un état et la présence d'un trait est ensuite normalisé par la durée du passage par cet état pour calculer la probabilité conjointe de chaque macro-trait et de l'état du modèle. Cette information est ensuite stockée avec les paramètres de chaque modèle sous la forme de cette troisième matrice de probabilité introduite plus haut.

Afin d'expliquer la méthode retenue pour exploiter ces informations, il est nécessaire d'observer ce qui se passe au niveau de l'évolution des probabilités cumulées de chaque état lorsque le modèle ne correspond pas au mot prononcé. On peut faire figurer ces évolutions sur un graphique en représentant à tout instant la probabilité de chaque état par un cercle de rayon plus ou moins grand, et les transitions par un trait reliant deux états à deux instant consécutifs.

La figure 2 décrit ainsi le comportement des modèles *déplace*, *détruis* et *dessine* quand le mot prononcé est *déplace*. Le signal et les graphes sont synchrones, ces derniers décrivant de gauche à droite l'évolution des états classés verticalement du premier au cinquième. Le chemin marqué en gras est celui que l'on retrouve par *backtracking*.

Ce type de schéma fait ressortir les chemins erratiques empruntés par le mauvais modèle lors de la reconnaissance : ainsi le troisième état du modèle *détruis* est parcouru tout le long des phonèmes /e/ à /s/ alors que le mot *détruis* ne comportant pas de fricative voisée, il ne peut y avoir simultanéité de cet état et du phonème /s/. De même la plosive sourde /p/ durant laquelle le modèle *dessine* se trouve dans le premier état.

Ces deux macro-traits étant généralement bien reconnus par un DAP à règles, le système hybride utilisera ce dernier pour extraire cette information du signal et donc pénaliser certains états des HMMs selon les probabilités contenues dans la troisième matrice issue de l'apprentissage.

Le nombre d'états étant fixé à 5 par modèle quel que soit le nombre de phonèmes par mot, on peut se poser la question du bien fondé d'un nombre d'état fixe. On constate en effet sur les mots courts (deux ou trois phonèmes) une redondance de certains états (un seul phonème *couvrant* plusieurs états), ce qui conduit à un déséquilibre entre ceux-ci et des mots plus longs quand à la finesse de modélisation par les HMMs. Obtenir par le DAP une information sur le nombre de phonèmes composant le mot permet d'adapter le nombre d'états du modèle en conséquence.

De même en ce qui concerne les probabilités initiales qui sont habituellement calculées par distribution équi-répartie des symboles entre les états du modèle puis évaluation de leurs fréquences d'apparition, tandis que les probabilités de transition sont fixées à des valeurs empiriques. La connaissance des macro-traits, de leurs positions et du nombre d'état du modèle permet une évaluation plus fine de ces valeurs, ce qui est très intéressant quand on sait l'importance de celles-ci sur la convergence des modèles probabilistes.

Afin de ne pas s'écarter du but recherché, il faut prendre soin de ne pas dégrader les performances des HMMs bien sûr. C'est pourquoi une part importante des expérimentations effectuées cherche à déterminer dans quelle mesure les résultats du DAP doivent modifier les probabilités des HMMs pour faire chuter celles des mauvais modèles sans perturber le modèle correct. La simulation du DAP à règles autorisant un contrôle des taux de reconnaissance de chaque macro-trait, il est ainsi possible de fixer pour chacun d'eux le seuil en deça duquel la détection par les règles ne doit plus être utilisée pour contraindre les HMMs.

3 Résultats expérimentaux

Les expériences préliminaires ont été effectuées sur une partie du corpus d'*ICPplan* constituée de 14 mots prononcés 10 fois par 3 locuteurs masculins, soit 30 répétitions par mots dont 6 ont servi de corpus d'apprentissage. L'ensemble du corpus a été étiqueté phonétiquement à la main ce qui à permis, pour le corpus d'apprentissage, de fixer le nombre d'états à un par macro-trait, puis d'initialiser les matrices de probabilités d'émission et de transitions des modèles.

Suite à l'apprentissage par algorithme de Baum-Welch, les matrices des probabilités conjointes des macro-traits et des états ont été calculées par l'algorithme de Viterbi suivi d'un *backtracking* tel que

décrit au paragraphe 2.

La simulation du DAP à base de règles a consisté à transformer les étiquettes phonétiques de l'étiquetage manuel en macro-traits sous la condition qu'un nombre pseudo aléatoire dépasse un seuil fixé pour chaque macro-trait en se basant sur les performances du système *DIRA*. En cas de non dépassement du seuil, aucune indication n'est fournie au modèle HMM contraint.

Dans le cas contraire, lorsque le macro-trait est annoncé, une procédure insérée dans la boucle de parcours du temps de l'algorithme de Viterbi consulte la troisième matrice pour chaque état. Si le trait prédit possède une probabilité inférieure à un seuil, la probabilité de l'état est multiplié par une pénalisation.

On le voit ici, les paramètres influençant le modèle hybride sont donc le taux de reconnaissance du DAP à règles pour chaque macro-trait, le seuil en deç'a duquel on pénalise l'état, et la valeur de la pénalisation elle même. Si les premiers sont issus des performances de *DIRA*, on a choisi de fixer le seuil minimal de probabilité d'un macro-trait à 1%, et la pénalisation est la probabilité elle même, plus une valeur minimale, pour des raisons de dépassement de capacités.

Avec ces valeurs, le taux de reconnaissance en mots isolés passe de 95% pour les modèles HMM classiques à 97%, ce qui est déjà un résultat encourageant si l'on considère que les paramètres ont été déterminés par tâtonnements. Il est clair que l'on doit obtenir un taux de quasiment 100% sur un si petit corpus, et les expériences menées actuellement tendent vers ce but. Cependant, il est intéressant de constater que, par rapport au modèle classique, les erreurs éliminées sont bien celles qui ont motivé l'exploration de cette méthode d'hybridation, à savoir des confusions entre mots phonétiquements très éloignés tels que *blanc* et *déplace*.

Ainsi sur les dix mots du corpus de test encore confondus, le modèle ayant la plus forte probabilité est toujours celui d'un mot possédant les mêmes macro-traits et dans le même ordre, au succès du DAP près : l'article *le* par exemple est encore parfois reconnu comme *bleu*. Ce genre de défaut pourrait sans doute être corrigé en ne considérant pas seulement les macro-trait détecté, mais également ceux qui ne le sont pas, en fonction de la performance du DAP à les prédire.

4 Discussions

En comparaison avec d'autres systèmes mettant en jeu des modules déterministes et probabilistes tels que décrits dans [5] ou [4] où les modèles probabilistes vérifient les hypothèses induites par les règles en alignant des formes correspondant à des éléments phonologiques , il s'agit bien ici de modifier le comportement des HMMs de façon dynamique.

Une manière de se rapprocher de ces méthodes dans le contexte qui nous intéresse peut être une présélection du vocabulaire parmi les mots du corpus après un premier étiquetage grossier. Le système SUMMIT [12] serait plus proche avec cependant comme différence principale que les symboles alignés par algorithme de Viterbi sont dans ce cas directement les phonèmes détectés par règles, et non des symboles calculés par quantification vectorielle.

Les résultats présentés n'ont bien sûr pas pour l'instant une valeur statistique suffisante. Mais avant d'effectuer les tests sur des corpus de l'ordre de plusieurs centaines de mots (il s'agit toujours d'un système orienté *commande vocale*), il convient d'affiner le modèle, en particulier quant à l'influence des différents paramètres. De même il est question d'offrir la possibilité de modifier plus profondément la topologie des HMMs avant apprentissage grâce à l'analyse du DAP, par exemple en autorisant le saut d'un ou plusieurs états quand cela est pertinent, comme dans le cas d'une voyelle muette en fin de mot. Enfin les contraintes au niveau des transitions entre modèles dans le cas des mots connectés devront être également étudiées.

5 Conclusion

Les bases d'un modèle hybride pour la reconnaissance automatique de mots enchaînés ont été posées. Il tente de tirer partie des avantages des méthodes globales et analytiques de décodage acoustico-phonétique tout en conservant une certaine simplicité. Les premiers résultats qui sont engageants ont permis de mettre à jour de nouvelles possibilités d'hybridation. La suite des travaux va consister à tester ces options et à valider le système de façon plus déterminante, et en particulier par l'intégration d'un DAP à règles réel dans un système opérationnel complet.

Références

[1] Marie-Luce Bourguet. "ICPplan: Dialogue multimodal pour la conception de plans architecturaux". *Actes 19ème JEP*, pages 369–374, 1992.

[2] J.S. Bridle, M.D. Brown, and R.M. Chamberlain. "Continuous Connected Word Recognition Using Whole Word Templates". *The Radio and Electronic Engineer*, 53 :167–175, 1983.

[3] Jean Caelen, Emanuel Reynier, Philippe Verdier, and Alberto Lichene. "Reconnaissance de la parole dans le projet MultiWorks". *Actes 19ème JEP*, pages 249–254, 1992.

[4] L. Deng, V. Gupta, M. Lenning, P. Kenny, and P. Mermelstein. "Acoustic Recognition Component of an 86000-words Speech Recognizer". *ICASSP*, pages 741–744, 1990.

[5] A Farhat, G Perennou, and N Vigouroux. "Segmentation en événements phonétiques et modèles markoviens HMM pour l'étiquetage phonotypique". *Actes 19ème JEP*, pages 297–302, 1992.

[6] J.P Haton, N Carbonel, D. Fohr, J.F. Mari, and A. Kriouille. "Interaction Between Stochastic Modeling and Knowledge-Based Techniques in Acoustic Phonetic Decoding of Speech". *ICASSP*, 2 :868–871, 1987.

[7] Denis Jouvet and Christian Gagnoulet. "Développements récents en reconnaissance de la parole". *L'Echo des Recherches*, 135 :27–36, 1989.

[8] Chin-Hui Lee and Lawrence R Rabiner. "A Network-Based Frame Synchronuous Level Building Algorithm for Connected Word Recognition". *ICCASP*, 1 :410–413, 1988.

[9] Lawrence R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proc. of the IEEE*, 77(2) :257–286, 1989.

[10] Hélène Tattegrain. "Un système expert de décodage acoustico-phonétique de la parole continue". Thèse de l'Institut National Polytechnique de Grenoble, 1990.

[11] Hayan Ye. "The Final Report of MultiWorks". Technical report, ICP Grenoble, 1990.

[12] V. Zue, J. Glass, M. Phillips, and S. Seneff. "the MIT SUMMIT Speech Recognition System : a Progress Report". *Proc. of the 2nd DARPA Speech and Language Workshop*, 1989.

XX èmes

JOURNEES D'ETUDE
SUR LA PAROLE

SESSION ORALE : O7

PROSODIE ET PSYCHO-PERCEPTION

ACCENT, DUREES SYLLABIQUES ET SEGMENTATION LEXICALE

Marie-Hélène Banel et Nicole Bacri

Laboratoire de Psychologie Expérimentale, Université René Descartes, URA 316 CNRS, Paris, France

RESUME
La pertinence des patterns accentuels dans la segmentation lexicale est attestée en anglais. Mais la clarté des frontières syllabiques en français pourrait la rendre inopérante. Les expériences présentées ici font partie d'une série d'études sur la segmentation de mots disyllabiques qui comportent des monosyllabes enchassés. Les disyllabes présentent un pattern accentuel iambique ou trochaïque. Le pattern iambique entraine majoritairement l'identification des disyllabes, le pattern trochaïque leur segmentation, ce qui atteste la pertinence d'une stratégie de segmentation métrique en français (Expérience 1). Par contre, quand les syllabes sont de même durée, longue (Exp. 2) ou brève (Exp. 3), fréquence des monosyllabes et structure syllabique ont un effet conjoint sur la segmentation lexicale: la procédure de traitement est alors à base syllabique.
MOTS CLE: *perception, segmentation, accent, durées.*

La reconnaissance de la parole implique sa segmentation en mots. Les indices de segmentation ne sont pas toujours suffisants pour marquer les frontières de mots, particulièrement pour la parole continue. De nombreuses recherches ont analysé le rôle spécifique d'unités de taille phonémique ou syllabique dans la segmentation. Ces unités varient en fonction de la langue étudiée. Par ailleurs, il est apparu que la structure prosodique pourrait fournir les indices les plus fiables de la segmentation lexicale (Norris et Cutler, 1988).

La présente étude analyse la segmentation lexicale lorsqu'un mot disyllabique comporte deux noms monosyllabiques, afin d'évaluer le poids des procédures de segmentation basées soit sur les indices prosodiques, soit sur un accès direct aux unités constituant le disyllabe. Le pattern rythmique correspondant à un contraste temporel entre les deux monosyllabes contenus dans le mot disyllabique pourrait être utilisé au cours de la segmentation lexicale. En anglais, les syllabes accentuées sont plus faciles à traiter que les syllabes non accentuées: l'extraction d'un mot est plus facile quand il se situe au début d'un non-mot disyllabique non-accentué sur la dernière syllabe, que lorsque les deux syllabes sont accentuées (Cutler et Norris, 1988). L'alternance régulière de syllabes fortes et faibles peut permettre l'établissement d'une routine de segmentation perceptive (Cutler, 1986; Grosjean et Gee, 1987). Même si les différences de *lexical stress* n'affectent pas toujours directement l'accès au lexique (*trustee* et *trusty* sont traités comme des homophones, Cutler, 1986), une syllabe accentuée peut déclencher l'accès au lexique, y compris quand elle ne se situe pas en position initiale.

Comme Pitt et Samuel (1990) l'ont souligné, la régularité temporelle peut fournir "a stable structure that the processing system can capitalize on by predicting the future occurrence of perceptually clear segmental speech" (p. 564). Les auditeurs pourraient élaborer sur la base de cette régularité une stratégie de segmentation métrique (Cutler, 1990). Les données disponibles suggèrent fortement que la prosodie, et plus précisèment les patterns accentuels ont un rôle fonctionnel dans la segmentation de la parole. Cependant ces données ne sont pas concluantes en ce qui concerne le statut psychologique des patterns rythmiques dans le traitement perceptif. La durée en tant que telle pourrait suffire à rendre compte d'un accès plus facile aux monosyllabes plus longs. On peut au contraire estimer que les patterns rythmiques facilitent la segmentation en suscitant une stratégie décisionnelle spécifique (Pitt et al., 1990). Notre étude a pour but d'analyser la fonction perceptive des propriétés métriques en français, langue où l'accent est situé sur la dernière syllabe des mots plurisyllabiques. Cet accent étant prévisible, une stratégie de segmentation métrique devrait être plus efficace qu'en anglais en dépit de l'absence de valeur distinctive lexicale de l'accent.

Un groupe rythmique non emphatique en français est caractérisé par un fort allongement de la syllabe finale par rapport aux syllabes précédentes (Rossi, Di Cristo, Hirst, Martin et Nishinuma, 1981). Malgré la tendance à accentuer la syllabe initiale des mots dans certains styles oraux (par ex. le style radiophonique, Fonagy, 1980), la structure iambique (bref-long) apparait comme la structure rythmique de base en français (Hirst et Di Cristo, sous presse, Fletcher, 1991, Vaissière, 1991). Bien que

l'allongement final soit le plus souvent accompagné d'un mouvement de F0, il est en lui-même perceptivement saillant. L'allongement peut donc constituer un indice fiable d'identification de la frontière droite et servir à la segmentation lexicale (Wenk et Wioland, 1982).

Le rôle des groupes rythmiques dans les processus de segmentation de la parole a été testé en examinant si l'inversion du pattern accentuel de mots disyllabiques affecte la segmentation lexicale. Si l'on admet que le français est une langue *trailer-timed* (à allongement terminal) dont les groupes rythmiques n'excèdent généralement pas deux ou trois syllabes (Wenk et al., 1982) l'allongement final d'une séquence de deux syllabes devrait induire les auditeurs à assembler la première syllabe brève avec la seconde syllabe longue pour en dériver une seule unité lexicale cohérente. La comparaison des performances de segmentation lexicale en fonction de la présence du pattern rythmique iambique usuel en français ou du pattern inversé, trochaïque, permettra de déterminer si les groupes rythmiques jouent un rôle dans la segmentation de séquences disyllabiques en leurs monosyllabes en fonction du pattern accentuel. Selon l'hypothèse d'une segmentation basée sur le rythme, un pattern iambique entrainera moins de segmentations qu'un pattern trochaïque.

Cependant, les données disponibles sur l'unité de base dans la segmentation des langues romanes, la syllabe, ne sont pas en faveur de cette hypothèse. La rapidité de la détection de syllabes dépend de paramètres structuraux et non pas temporels (Mehler, Dommergues, Frauenfelder et Segui, 1981; Dupoux et Mehler, 1990). Un monosyllabe en position initiale peut déclencher l'accès au lexique quelle que soit sa durée. Ce mode d'accès syllabique semble plus pertinent qu'une stratégie de segmentation fondée sur l'accent pour des langues comme les langues romanes dont les frontières syllabiques sont claires. En catalan, les groupes rythmiques ne jouent pas de rôle dans l'accès au lexique: le changement de position de l'accent de disyllabes n'a pas d'effet sur la syllabation (Sebastián-Gallés, Dupoux, Segui et Mehler, 1992). La *transparence acoustique* des frontières syllabiques annule la fonction potentielle de la régularité temporelle. Dans cette optique, le pattern accentuel, normal ou inversé, ne devrait pas avoir d'effet sur la segmentation lexicale. Le traitement des propriétés métriques ne serait que post-lexical. Des procédures différentes s'appliqueraient aux monosyllabes, dont la fréquence élevée faciliterait l'accès, et aux disyllabes, dont le traitement serait surtout fonction de la complexité syllabique (Segui, 1984).

Les résultats d'une expérience portant sur la segmentation en leurs composants monosyllabiques de disyllabes présentant un pattern accentuel iambique ou trochaïque n'ont toutefois pas validé les prédictions dérivées de cette hypothèse "syllabique". Ils ont certes confirmé que la durée intrinsèque des syllabes n'avait pas d'effet sur l'identification des monosyllabes. Mais,- et c'est là le point essentiel-, la décomposition lexicale a été significativement accrue par un pattern trochaïque (46% de segmentations), alors qu'un pattern iambique suscitait l'identification des disyllabes (27% de segmentations). De plus, le traitement des seuls items présentant un pattern trochaïque fut sensible à la fréquence des monosyllabes le composant. Ces données suggèrent que les auditeurs ont utilisé les propriétés métriques des stimuli pour les segmenter et ne se sont pas appuyés sur une analyse syllabique (Bacri et Banel, 1993). Mais la faible sensibilité des temps de réponse aux caractéristiques prosodiques des stimuli laissait penser que la métrique avait pu influencer les décisions des sujets plus que leur traitement des mots. C'est afin de préciser les relations temporelles entre le traitement des patterns accentuels et la segmentation lexicale que les expériences présentées ici ont été réalisées, en focalisant l'attention sur les propriétés métriques des stimuli.

EXPERIENCE 1

La focalisation de l'attention des sujets sur les caractéristiques temporelles des stimuli devrait rendre la tâche plus aisée, diminuer les temps de réponse et faire apparaitre sur ces temps l'effet facilitateur du pattern iambique sur l'identification du disyllabe, celui du pattern trochaïque sur la segmentation, si le rythme est traité précocement. S'il n'intervient que tardivement dans la décision, ce sont les paramètres fréquentiels et structuraux qui devraient influencer les réponses, indépendamment du rythme.

Méthode:

Sujets: 10 étudiants de l'Université Paris V, de langue maternelle française.

Matériel: Les stimuli test (48 mots disyllabiques) comprenaient deux mots monosyllabiques (*marmotte: mare, motte*). Des 96 séquences disyllabiques de remplissage 80% étaient des suites de monosyllabes formant un non-mot, 20% des mots non décomposables en monosyllabes. Tous les items ont été enregistrés par un locuteur parisien de langue maternelle française. Ils ont été produits à débit constant, et la F0 a été maintenue aussi plate que possible. Deux patterns rythmiques ont été ensuite réalisés pour chaque disyllabe test: un pattern bref-long iambique, et un pattern long-bref trochaïque.

Chacune des deux versions a été construite par la concaténation de deux monosyllabes. Ainsi les items ne contenaient pas d'indices de coarticulation intersyllabique. La différence de durée entre les versions d'une même syllabe était d'environ 35% (durées de l'ordre de 550 ms vs. 350 ms). L'intensité a été égalisée. Les items de remplissage ont également été réalisés par la concaténation de deux syllabes, toutes les deux longues (environ 550 ms). Les mots test ont été choisis après deux expériences de décision lexicale afin de s'assurer que les mots rares appartenaient au lexique des sujets.

La fréquence des mots disyllabiques était moyenne-faible. Les monosyllabes enchâssés étaient rares ou fréquents. 4 combinaisons de fréquence des monosyllabes ont été réalisées: RR, RF, FR, FF (R: rare, F: fréquent). La structure syllabique de chaque séquence disyllabique a été contrôlée de façon à ce que 12 séquences disyllabiques (4 combinaisons de fréquence x 3 items) représentent chaque combinaison: CV-CV, CV-CVC, CVC-CV, CVC-CVC (C: consonne, V: voyelle).

Procédure

Chaque sujet a entendu les mots test et les items de remplissage (l'ordre de présentation était contrebalancé); l'intervalle entre deux stimuli successifs était de 3 s. Les sujets devaient répondre le plus vite possible s'ils avaient entendu un mot disyllabique ou plusieurs mots. La consigne insistait particulièrement sur le fait qu'ils allaient entendre des séquences de deux syllabes, afin de diriger indirectement leur attention sur les propriétés structurales des stimuli. Les temps de réponse (TR) ont été mesurés à partir de la fin des séquences.

Résultats et discussion

Les pourcentages de segmentation en fonction du rythme, de la fréquence des monosyllabes et de la structure syllabique sont indiqués Figure 1. Le rythme iambique a entrainé plus d'identifications du disyllabe que le rythme inversé (73% vs.62%). Une analyse de variance portant sur le nombre de réponses de segmentation a montré que le rythme ($F(1, 9)=24.3$, $p<.001$), la fréquence ($F(3, 27)=5.21$, $p<.01$) et la structure syllabique ($F(3, 27)=13.3$, $p<.001$) avaient eu un effet significatif, mais non l'interaction entre le rythme et la fréquence (Figure 2). La fréquence n'a eu d'effet significatif que pour le rythme trochaïque: un monosyllabe fréquent en position finale a entrainé plus de segmentations qu' un monosyllabe rare (Scheffé, $p<.005$).

Les TR ont été plus rapides pour le pattern iambique que pour le pattern trochaïque (TR moyen: 786 ms (sd=495) vs. 911 ms (sd=459)). Pour chaque condition de fréquence des monosyllabes les TR moyens ont été: RR: 849 ms (sd=470), RF: 896 ms (sd=459), FR: 818 ms (sd=500), FF: 830 ms (sd=488). Les TR moyens ont varié en fonction de la structure syllabique: CV-CV: 921 ms (sd=488), CV-CVC: 793 ms (sd=500), CVC-CV: 849 ms (sd=459), CVC-CVC: 831 ms (sd=470). Les effets du rythme ($F(1, 9)=36.1$, $p<.005$) et de la structure syllabique ($F(3,27)=3.76$, $p<.05$) ont été significatifs, mais non l'effet de la fréquence ou les interactions entre facteurs. Le pattern iambique a conduit à une identification du disyllabe plus rapide que le pattern trochaïque (TR moyen: 669 ms (sd=504) vs. 848 ms (sd=459), $z=2.49$, $p<.014$). Pour les deux rythmes, les identifications du disyllabe ont été plus rapides que les segmentations (pattern iambique: $z=3.35$, $p<.001$; pattern trochaïque: $z=2.26$, $p<.024$). On note l'absence de corrélation entre la longueur syllabique et les TR ($r^2=.0031$).

Figure 1(expérience1): pourcentage de segmentations pour chaque condition de rythme, de fréquence et de structure syllabique:

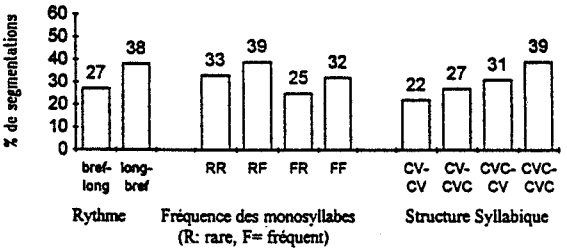
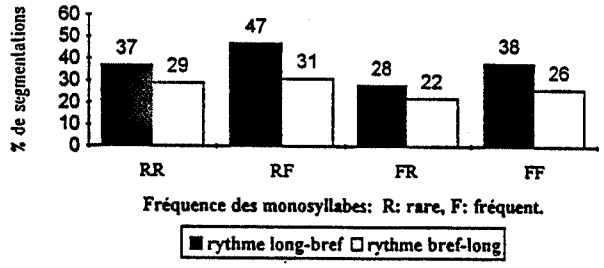


Figure 2 (expérience 1): pourcentage de segmentations en fonction du rythme et de la fréquence.



La focalisation indirecte de l'attention sur les propriétés métriques des stimuli test a mis en évidence l'effet du rythme, alors même que l'attention des auditeurs était conjointement attirée par la consigne sur les monosyllabes enchâssés. Comme lors de nos expériences antérieures dans lesquelles l'attention n'était pas focalisée (Bacri et Banel, 1993), le pattern iambique usuel continue à entrainer majoritairement la reconnaissance du disyllabe, alors que le pattern trochaïque accroît significativement le nombre de segmentations. L'hypothèse que les sujets ont développé une stratégie de segmentation métrique est donc confirmée, d'autant plus que

l'allègement de la tâche permet à cette stratégie d'être réellement opérationnelle: en dépit d'une dispersion importante des temps de réponse, le pattern iambique entraine une identification des disyllabes clairement plus rapide que le pattern trochaïque. On peut considérer que le pattern trochaïque inhibe le traitement de la séquence disyllabique.

Le fait que, en ce dernier cas seulement, se manifeste un effet facilitateur d'une fréquence élevée du monosyllabe en position finale, et de sa structure (les mots CVC sont plus souvent et plus rapidement reconnus que les mots CV), montre cependant que les auditeurs ont été sensibles aux variations des paramètres structuraux. C'est l'effet de ces paramètres que les expériences suivantes ont pour but d'étudier dans des conditions où le rythme lexical sera neutralisé par l'égalisation des durées des monosyllabes composant le disyllabe.

EXPERIENCE 2

Dans cette deuxième expérience, tous les stimuli (test et remplissage) sont des spondées (ils présentent un pattern long-long). Une procédure de segmentation prenant appui sur la syllabe devrait être plus efficace que précédemment et se marquer par une forte interaction entre la fréquence des monosyllabes et la structure syllabique: un monosyllabe devrait être accessible d'autant plus rapidement qu'il sera de type CVC. De plus, et dans l'hypothèse que les auditeurs ont un accès direct aux monosyllabes, mais doivent procéder à une analyse prélexicale pour accéder aux disyllabes (Segui, 1984), le nombre de segmentations devrait augmenter, et leurs TR être rapides. Si l'absence de patterns accentuels contrastés conduit les sujets à modifier leur mode de traitement dans la contre-épreuve que constituent cette expérience et la suivante, nous considérerons que notre hypothèse d'un traitement de la prosodie lors d'une étape précoce de l'accès au lexique sera renforcée.

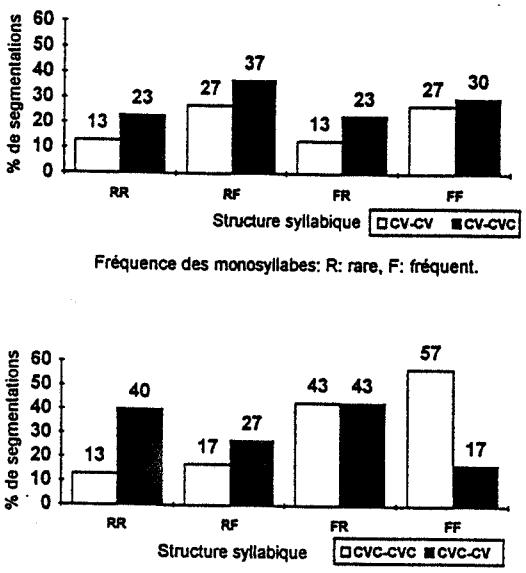
Méthode

L'expérience 2 est construite avec les mêmes items et selon les mêmes modalités que l'expérience 1, à la seule différence que tous les items ont été construits par concaténation de deux syllabes également longues (durée moyenne 500 ms). Dix nouveaux sujets ont passé cette expérience.

Résultats et discussion

D'une façon générale, le nombre de segmentations (28.5%) demeure proche de celui de l'expérience 1 (27%). La Figure 3 montre le pourcentage de segmentations en fonction de la fréquence et de la structure syllabique des monosyllabes. On constate qu'une fréquence élevée des monosyllabes et la présence d'une syllabe CVC accroissent les segmentations. Une analyse de variance montre toutefois que seule l'interaction entre ces facteurs ($F(9, 81)=3.86$, $p<.001$) est significative, et non leur effet principal. Ce résultat est dû à ce qu'une syllabe CVC de fréquence élevée induit plus de segmentations, quelle que soit sa position, qu'un monosyllabe CVC rare, alors que la fréquence d'un mot CV n'a pas d'effet sur les réponses.

Figure 3 (expérience 2): pourcentage de segmentations en fonction de la fréquence et de la structure syllabique:



Fréquence des monosyllabes: R: rare, F: fréquent.

L'analyse des TR conduit à des résultats analogues: absence d'effet principal de la fréquence et de la structure syllabique, interaction significative entre ces facteurs ($F(9, 81)=3.01$, $p<.01$). Un monosyllabe de fréquence élevée en position initiale ralentit les TR, lorsque le second monosyllabe (fréquent ou rare) est de même structure syllabique que le premier monosyllabe. Les séquences CV-CV sont traitées en moyenne en 907 ms (sd=380 ms) si la première syllabe est un mot fréquent, en 784 ms si c'est un mot rare (Scheffé, $p<.05$), les séquences CVC-CVC en 884 ms dans le premier cas, en 768 ms dans le second (Scheffé, $p<.05$). Dans tous les cas, l'identification des disyllabes est plus rapide que les segmentations (753 ms vs. 1061 ms), et les TR ne sont pas corrélés à la durée des mots ($r^2=.003$).

On notera tout d'abord que, contrairement à ce que pouvait laisser prédire l'hypothèse de voies d'accès distinctes aux monosyllabes et aux disyllabes, l'égalisation des durées et la focalisation de l'attention sur la présence des monosyllabes n'ont entrainé ni

une augmentation du nombre global de segmentations, ni l'accélération des réponses de ce type. Mais l'interaction entre fréquence et structure syllabique montre que les sujets ont effectivement extrait du signal les unités syllabiques et construit des représentations syllabiques. Ils ont traité les régularités phonotactiques et segmenté plus souvent la séquence quand une consonne finale marquait une frontière syllabique claire. Toutefois il est notable que la présence d'un monosyllabe fréquent au début de la séquence (CVC-CV), loin de faciliter un accès direct au lexique, n'a pas accru le nombre de segmentations et a ralenti la décision. L'analyse des disyllabes en leurs composantes a pris un temps suffisamment long pour que, dans la compétition entre hypothèses lexicales, la sortie la plus rapide ait été disyllabique. L'expérience suivante a pour but d'étudier si la procédure de traitement à base syllabique, qui vient de se manifester en l'absence de patterns accentuels contrastés, est sensible à la durée des syllabes et se modifie quand la pression temporelle augmente sous l'effet d'une réduction de la durée des stimuli.

EXPERIENCE 3

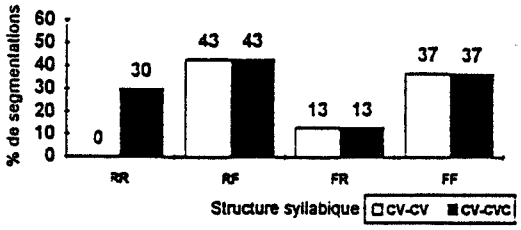
Méthode

L'expérience 3 est en tous points semblable à l'expérience 2, sous la réserve que les syllabes ont été compressées à 35% de leur valeur d'origine. Dix nouveaux sujets ont passé l'expérience.

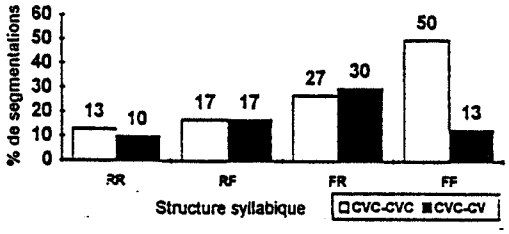
Résultats et discussion

Le nombre global de segmentations est demeuré stable (25%). L'effet de la fréquence des monosyllabes s'est avéré significatif ($F(3, 27)=6.64$, $p<.01$). Une séquence de mots rares n'entraine que 13% de segmentations, une séquence de deux mots fréquents en entraine 34% (Scheffé, $p<.01$). La présence d'un mot fréquent en position finale conduit à segmenter plus que s'il est en position initiale (30% vs. 21%, Scheffé, $p<.01$) (Figure 4). La significativité de l'interaction entre fréquence et structure syllabique témoigne à nouveau en faveur d'une procédure de traitement à base syllabique ($F(9, 81)=3.94$, $p<.001$). Contrairement aux conclusions que Dupoux et al. (1990) avaient tirées d'une tâche de détection de phonème, la durée des items a un effet sur l'accès au lexique dans une tâche d'identification de mots: seuls les monosyllabes fréquents sont assez rapidement activés pour déclencher un accés lorsque leur durée est de l'ordre de 350 ms. La structure syllabique a été analysée par les auditeurs: un mot fréquent en position finale induit d'autant plus de segmentations qu'il est de type CVC (Scheffé, $p<.05$). Toutefois, la brièveté des stimuli a accru la difficulté de la tâche au point que l'étude des TR montre des valeurs peu différenciées, de l'ordre de 900 ms.

Figure 4 (expérience 3): pourcentage de segmentations en fonction de la fréquence et de la structure syllabique:



Fréquence des monosyllabes: R: rare, F: fréquent.



DISCUSSION GENERALE

La tâche de segmentation de mots disyllabiques en leurs composants monosyllabiques est sensible à l'organisation des durées syllabiques. La structure du monosyllabe, en particulier la présence d'une consonne finale (CVC- et surtout -CVC), n'est utilisée directement comme indice de segmentation que lorsqu'un pattern accentuel trochaïque vient perturber l'application d'une stratégie à base métrique. Quand, dans les expériences 2 et 3, les durées des syllabes successives sont du même ordre, longues ou brèves, une structure CVC n'entraine plus de segmentations que si elle correspond à un mot de fréquence élevée, - effet restreint à un mot de ce type en position finale lorque les durées sont compressées. La décomposition s'appuie alors sur une analyse conjointe des dimensions structurales et lexicales. Cette analyse prend du temps. Les segmentations sont systématiquement plus longues que l'identification du seul mot disyllabique. La complexité syllabique accroit le nombre de segmentations, mais aussi les temps de réponse. L'activation lexicale n'est ainsi pas réellement facilitée par l'analyse structurale. Ces résultats suggèrent que les auditeurs ont bien adopté une procédure de segmentation à base syllabique, mais ils ne l'ont pas utilisée pour un accès direct au

lexique. Ce point est important: l'unité de segmentation semble devoir être dissociée de l'unité d'accès.

Le résultat essentiel de cette série d'expériences réside dans la confirmation d'un effet massif des patterns accentuels sur la segmentation. Cet effet confirme l'hypothèse que plus un pattern accentuel est prévisible, et plus le système perceptif l'utilisera pour anticiper les frontières de mots (Cutler, 1990; Pitt et al., 1990). L'application d'une stratégie de segmentation métrique est indépendante de la durée intrinsèque des segments, comme le montre l'absence de corrélation entre segmentations et durées syllabiques. Elle est donc probablement indépendante du débit. Elle conduit à percevoir spontanément une séquence de deux syllabes comme un seul mot, même en l'absence d'indices de co-articulation intersyllabiques. La présence du pattern iambique usuel semble rendre inopérant le traitement de la fréquence des mots ou de leur dimension structurale. Ce n'est que lorsque cette stratégie à base métrique ne peut s'appliquer que d'autres types de traitement organisent le comportement, ce qui suggère que le traitement de la prosodie intervient lors d'une étape précoce du traitement du signal de parole.

Remerciements
Cette recherche a été partiellement financée par le G.D.R. 957 "Sciences Cognitives de Paris" (C.N.R.S.).

Références

Bacri, N, & Banel, M-H, (1993). Rhythmic patterns and lexical parsing in French, *ESCA Workshop on Prosody,* Lund University Working Papers, 41, 120-123.

Cutler, A (1986). Forbear is a homophone: lexical prosody does not constraint lexical access, *Language and Speech*, 29, 201-220.

Cutler, A (1990). Exploiting prosodic probabilities in speech segmentation, in G. T. M. Altman (ed), *Cognitive models of speech processing: psycholinguistic and computational perspective.* (pp. 105-121). Cambridge: M.I.T. press.

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.

Dupoux, E, & Mehler, J (1990). Monitoring the lexicon with normal and compressed speech: frequency effects and the prelexical code, *Journal of Memory and Language*, 29, 316-335.

Fletcher, J (1991). Rhythm and final lengthening in French, *Journal of Phonetics*, 19, 193-212.

Fonàgy, I (1980). L'accent français: accent probabilitaire, in Fonàgy, I, & Léon, P. R (eds), L'accent en français contemporain. *Studia Phonetica*, 15, 123-233.

Grosjean, F, & Gee, J (1987). Prosodic structure and spoken word recognition, *Cognition*, 25, 135-155.

Hirst, D. J & Di Cristo, A. (sous presse). *Intonation Systems. A survey of twenty languages.* Cambridge University press.

Mehler, J, Dommergues, J. Y, Frauenfelder, U, & Segui, J (1981). The syllable's role in speech segmentation, *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305.

Norris, D, & Cutler, A (1988). The relative accessiblity of phonemes and syllables, *Perception and Psychophysics*, 43, 541-550.

Pitt, M. A, & Samuel, A. G (1990). The use of rhythm in attending to speech, *Journal of Experimental Psychology: Human Perception and Performance*, 16 (3), 564-573.

Rossi, M, Di Cristo, A, Hirst, D. J, Martin, Ph, & Nishinuma, Y (1981). *L'intonation. De l'acoustique à la sémantique.* Paris, Klincksieck.

Sebastiàn-Gallés, N, Dupoux, E, Segui, J, & Mehler, J (1992). Contrasting syllabic effects in Catalan and Spanish, *Journal of Memory and Language*, 31, 18-32.

Segui, J (1984). The syllable: a basic perceptual unit in speech processing?, in Bouma, H, & Bouwhuis, D. G (eds), *Attention and Performance X: Control of Language Processes*. (pp. 165-181). Hillsdale, N.J.: Erlbaum.

Vaissière, J (1991). Language independent features, in Cutler, A, & Ladd (eds), *Prosody models and measurements.* (pp. 53-66). Springer-Verlag.

Wenk, B, & Wioland, F (1982). Is French really syllable-timed?, *Journal of Phonetics*, 10, 193-216.

'LECON' DE 'DECLINAISON' DE MARGUERITE DURAS ET DE MARGUERITE YOURCENAR

Zsuzsanna FAGYAL

Institut de Linguistique de l'Académie des Sciences de Hongrie
H-1250 Budapest BP. 19, Hongrie
Institut de Phonétique de l'Université de Paris III
CNRS-UA 1027, 19, rue des Bernardins 75005 Paris, France

Résumé

Cette étude se propose d'analyser les variations mélodiques globales dans deux extraits de parole d'interview en terme de 'déclinaison' de la fréquence fondamentale. L'indice de 'déclinaison', défini comme la pente de la droite de régression tracée entre les valeurs maximales de F0 en fonction du rang de la syllabe, est calculé dans les groupes de souffle qui remplissent les conditions suivantes: (1) début d'un nouveau 'paragraphe' dans le discours, (2) contour terminal descendant, (3) nombre de syllabes allant de 13 à 18, (4) durée moyenne comprise entre 3 et 4 secondes, (5) débit moyen entre 5 et 6 syll/sec. Suit une discussion sur la 'déclinaison' de F0 dans la parole non-lue et sur les différences dans la gestion du contour mélodique global entre les deux locutrices: M. Duras et M. Yourcenar.

Mots-clés: *parole non-lue, déclinaison, paragraphe*

INTRODUCTION

La baisse progressive de la fréquence fondamentale au cours d'une unité intonative ('déclinaison de F0') est considérée comme un phénomène universel non seulement dans la communication humaine, mais également chez les primates [Hauser-Fowler, 1992]. Elle est, en partie, physiologique: après inspiration de l'air précédant la phonation, la pression sous-glottique - l'un des paramètres qui influe sur la fréquence laryngienne - tend à diminuer progressivement au cours de l'expiration [Lieberman, 1967].

A la base d'exemples de parole lue, on démontre que la 'déclinaison' est aussi sous le contrôle du locuteur (entre autres [Gelfer et al., 1983]), et qu'elle sert d'indice de segmentation en structures grammaticales (entre autres [Pierrehumbert, 1979]).

La situation s'avère, néanmoins, plus compliquée que ne suggèrent les études menées majoritairement sur des phrases isolées. Même au prix de la restriction de la théorie de la 'déclinaison' sur les 'groupes de souffle' à contours terminaux descendants (entre autres [Lieberman et al., 1985]), les nombreux contre-exemples relevés surtout dans la parole spontanée demeurent problématiques. En l'absence de 'déclinaison' sur des phrases découpées d'une conversation spontanée selon le critère de la "bonne formation", Umeda remet en cause l'existence du phénomène dans la parole spontanée [Umeda, 1982].

Selon notre hypothèse, la 'déclinaison' existe également en parole non-lue, mais elle est plus fonction de la présence d'une argumentation cohérente que de modes de production. Lorsqu'un discours non-lu, tel notre corpus d'interview, est basé sur une argumentation élaborée ('well-planned'), il peut être segmenté en macro-unités discursives dont les frontières seront bien indiquées par des variations prosodiques (voir [Bruce, 1982] et [Thorsen, 1985] sur les fonctions textuelles de la prosodie). Notre hypothèse veut que le discours oral soit segmenté plutôt en unités argumentatives qu'en phrases isolées, concept non-opératoire hors du laboratoire.

Dans la présente contribution, nous nous proposons de montrer que: (1) à la base de la segmentation d'un discours d'abord en 'prises de parole', ensuite en 'paragraphe', puis en 'groupes de souffle', il est possible de mesurer la 'déclinaison' de F0 dans la parole non-lue (2) étant un phénomène réel, la 'déclinaison' peut servir d'indice mélodique global.

2. CORPUS ET MESURES

Le corpus est constitué de deux extraits d'interview télévisée de 10 minutes, enregistrée à 5 ans d'intervalle dans le cadre de la même émission ('Apostrophes' de Bernard Pivot) avec deux femmes écrivains françaises, Marguerite Duras (D) et Marguerite Yourcenar (Y) vers la fin de leur carrière professionnelle. Bien que les locutrices aient été enregistrées en direct, en raison de la présence des caméras, leur discours est particulièrement surveillé.

Les performances des deux locutrices interrogées par le même journaliste dans les mêmes conditions sont comparables du point de vue situationnel (entretien médiatisé), social (sexe, âge, profession) et psychologique (position de dominance [Duez, 1991]).

Afin d'obtenir une visualisation graphique des variations mélodiques globales dans de longs extraits du discours, la fréquence fondamentale a d'abord été détectée par un moyen analogique, et représentée ensuite sur des tracés de mingographe (échelle mélodique calibrée en quarts de ton; vitesse de défilement = 50 mm/s). Une détection numérique du type FFT du signal échantilloné à 16 kHz (logiciel UNICE, LIMSI CNRS) a servi de

base pour les mesures de F0. L'analyse acoustique a eu lieu au laboratoire de l'Institut de Phonétique de l'Université de Paris III.

3. SEGMENTATION DU DISCOURS

Plutôt que de suivre les études antérieures dans leur démarche de découper, à des endroits différents du corpus oral, des phrases "bien formées" selon les critères syntaxiques de l'écrit [Umeda, 1982] [Lieberman et al., 1985], nous proposons d'identifier ces unités comme le résultat d'un découpage en macro-unités: (1) prises de parole, (2) paragraphes, (3) groupes de souffles.

En raison de la nature dialogique du corpus, les *'prises de parole'* ('turns') des locutrices ont été considérées comme l'unité discursive et intonative la plus large. Les prises de parole d'une durée inférieure ou égale à 20 secondes (interjections, réponses courtes) ont été éliminées.

Les prises de parole ont ensuite été segmentées en *'paragraphes'* [Lehiste, 1975]. Le paragraphe dans la parole continue, comme à l'écrit, marque l'introduction d'un nouveau sujet ('topic change') dans le discours. Il a été prouvé pour de nombreuses langues que ces unités du discours possèdent certaines propriétés prosodiques qui permettent de les identifier même en absence d'informations segmentales (dernièrement [Swerts, 1991]). Pour l'identification des paragraphes à base d'indices prosodiques, le locuteur compare en premier lieu la tessiture mélodique des segments en cours à celle de leur contexte gauche immédiat [Kreiman, 1982]. Autrement dit, lorsque le rétrécissement de la plage des variations mélodiques sur un segment en cours (accompagné ou non d'autres indices de finalité, tels que l'allongement final, laryngalisation, pause) est immédiatement suivi de l'élargissement de la plage, voire de l'augmentation de la totalité du registre de F0 [Brown et al., 1980], le locuteur identifie l'apparition d'un nouveau paragraphe. Les figures 7 et 8 de l'annexe représentent deux extraits du corpus, segmentés selon l'ensemble de ces critères.

Les paragraphes ont été segmentés en unités plus petites à la base de la présence d'une pause, accompagnée ou non d'une reprise de souffle. Conformément à notre hypothèse, nous nous limitons aux séquences initiales des paragraphes, parce que leur fonction dans la structure argumentative est identique (introduction d'un nouveau sujet dans le discours). Dans le corpus D, la totalité de ces séquences (N=16) sont délimitées par une pause respiratoire. Dans le corpus Y, la majorité des séquences (93%, N=15) remplissent ce critère. Leurs occurences ont été identifiées à l'écoute par 5 musiciens français (voir aussi [Goldman-Eisler, 1968, p. 103]). Ceci nous a amenée à considérer ces séquences comme une unité et à les nommer *'groupe de souffle'*.

4. PLAGES MELODIQUES

Le premier groupe de souffle de chacun des paragraphes a été représenté sur une échelle logarithmique de F0 en fonction du 'rang' de la syllabe dans le groupe, indépendamment de leur longueur totale (figures 1 et 2). Chacune des syllabes est représentée par sa valeur maximale de F0. Cette méthode de superposition de contours, appelée 'procédure d'abstraction visuelle' (voir entre autres [Maeda, 1976]) est appliquée ici en vue de visualiser la plage des variations des valeurs de F0 chez les deux locutrices.

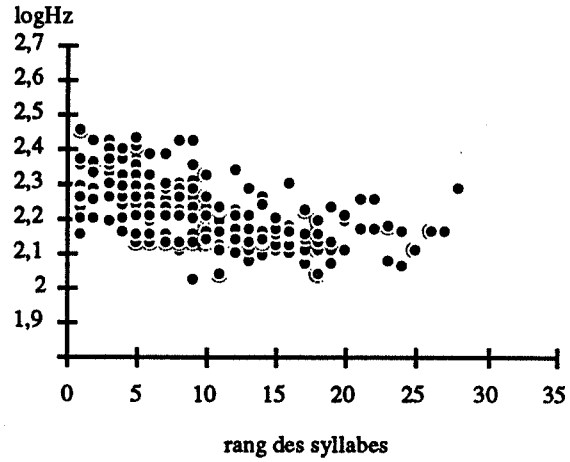


Figure 1. Valeurs maximales de F0 par syllabe dans la totalité des groupes initiaux de paragraphe (N=16) dans le corpus D.

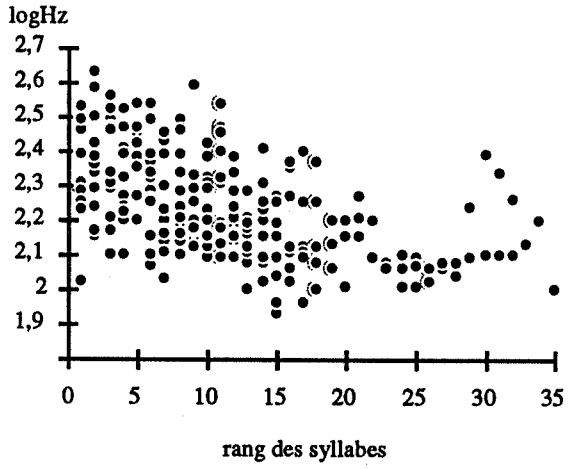


Figure 2. Valeurs maximales de F0 par syllabe dans la totalité des groupes initiaux de paragraphe (N=15) dans le corpus Y.

Il ressort des figures 1 et 2 que la fréquence fondamentale tend à baisser au fur et à mesure que l'on avance vers la fin des groupes. Il y a, néanmoins, des différences notables entre les deux locutrices. Il est visible que la plage mélodique de D (figure 1) est plus rétrécie que celle de Y (figure 2), et que le 'plateau' (ligne reliant les maximas de F0 [Vaissière, 1983]) est moins 'déchiqueté' chez D que chez Y. Ceci constituerait l'une des particularités globales du style vocal de la locutrice D [Fagyal, 1994].

Les différences entre les plages mélodiques peuvent, néanmoins, provenir de la longueur des groupes, ainsi

que des débits moyens de parole (tableau 1), sachant qu'un débit plus élevé (locutrice D) favorise le rétrécissement des intervalles de variations mélodiques [Deese, 1980]. L'indice de 'déclinaison' ne peut être calculé qu'après avoir fixé ces paramètres.

Afin de produire un contexte comparable, nous avons décidé de trier les groupes selon le maximum de paramètres, susceptibles d'influencer leur éventuelle tendance de 'déclinaison'.

	Groupes de souffle			
	N total	syll M	durée M	débit M
Locutrice			(s)	(syll/s)
D	16	17	2.93	5.8
Y	15	18,5	3.66	5

Tableau 1. (de gauche à droite): Nombre total des groupes de souffle superposés; nombre moyen des syllabes, durées moyennes et débits moyens pour la totalité des groupes (figure 1 et 2) chez les deux locutrices.

5. CONTROLE DES PARAMETRES DE LA DECLINAISON

Partant des travaux traitant sur la 'déclinaison', nous avons constitué la liste (non-exhaustive) des paramètres, susceptibles d'influencer la tendance avec laquelle la fréquence fondamentale décline ou non sur une unité intonative donnée (ici: 'groupe de souffle'). Ces paramètres sont: fonction argumentative du groupe dans le discours (apport d'information nouvelle), lieu de la reprise du souffle, direction de la pente du contour terminal, durée du groupe, débit, distribution des proéminences locales (accents de mots), influences segmentales au bas niveau (coarticulation).

La fonction argumentative des groupes (début d'un 'paragraphe') et le paramètre physiologique (reprise du souffle) sont considérés comme fixes conformément au protocole de segmentation en 'paragraphes' et en 'groupes de souffle'. Vue la variabilité lexicale inhérente à cette comparaison de parole non-lue, l'influence segmentale au bas niveau (consonnes, voyelles) ne peut être contrôlée. Ceci est une limitation de notre étude. Une partie des groupes de souffle des figures 1 et 2 a été sélectionnée par la suite selon les critères suivants: (1) contour terminal descendant, (2) nombre de syllabes allant de 13 à 18, (3) durée moyenne comprise entre 3 et 4 secondes, (4) débit moyen entre 5 et 6 syll/sec. Les variations de F0 à l'intérieur des groupes de souffle constituent notre variable dépendante.

Cinq groupes de souffle par locutrice ont rempli l'ensemble des conditions (leur nombre égal est dû au hasard). Ils sont représentés sur les figures 3 (D) et 4 (Y). Leur liste est donnée dans l'annexe.

Les indices de variations globales et les droites de régression linéaire ont été calculés (logiciel Excel 4.0). Il s'avère que les variations de F0 à l'intérieur des plages respectives sont différentes chez les deux locutrices ($F=1.33$; $p<0.01$). La moyenne des valeurs est significativement plus grande chez Y que chez D ('one-tailed independent t-test'; $t=1.65$; $p<0.01$).

La 'déclinaison' de F0 en fonction du rang de la syllabe dans le groupe de souffle a été exprimée en terme de droite de régression linéaire, tracée entre les valeurs de F0 selon la méthode du moindre écart.

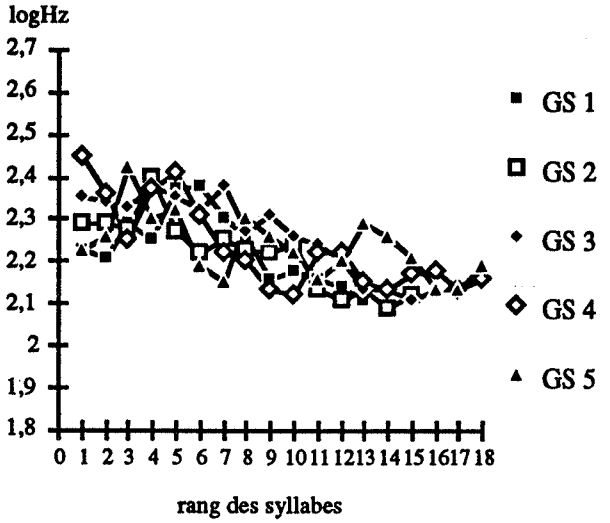


Figure 3. Valeurs maximales de F0 par syllabe dans cinq groupes initiaux de début de 'paragraphe' (GS) dans le corpus D (voir liste dans l'annexe).

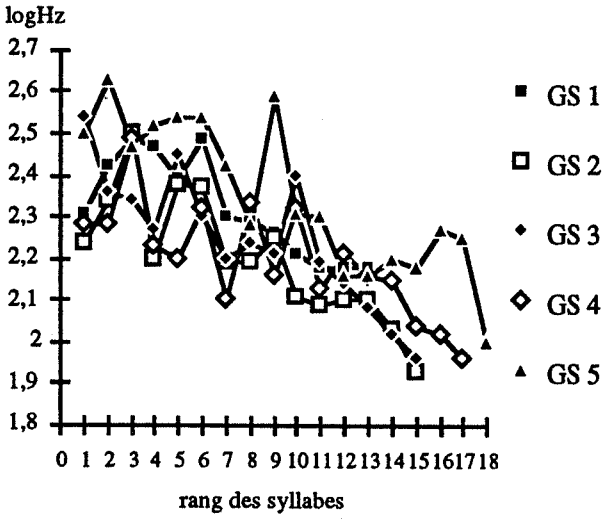


Figure 4. Valeurs maximales de F0 par syllabe dans cinq groupes initiaux de début de 'paragraphe' (GS) dans le corpus Y (voir liste dans l'annexe).

Le tableau 2 illustre les résultats des calculs de régression pour chacun des groupes de souffle des deux locutrices. On constate que la pente (légende 'a') de chacune des droites est négative, ce qui montre que les valeurs de F0 ont bien tendance à décroître au fur et à mesure que l'on avance vers les syllabes finales du

groupe. Il a y donc 'déclinaison'.

Cette tendance est significative ($p<0.05$) dans la totalité des groupes de souffle chez Y et dans quatre groupes sur cinq chez D. Autrement dit, l'évolution globale de F0 en fonction du rang de la syllabe dans le groupe de souffle peut être prédite en terme de 'déclinaison' sans risque d'erreur chez les deux locutrices.

Les valeurs d'interception (légende 'b'), correspondant aux rehaussements ('resettings') en début de paragraphe sont plus élevées chez Y. Ceci est lié aux différences entre l'étendue des plages. On note également qu'à durées de groupe et à débits comparables entre Y et D, les groupes de souffle de Y sont caractérisés par des pentes plus 'abruptes', dûes à des plus grands intervalles de variations mélodiques (σ_Y= 72,6 Hz; σ_D= 36,3 Hz). Nous allons essayer de démontrer cette tendance sur un exemple concret.

		a	r	b	dl	p
D				(Hz)		
	GS1	-4.9	-0.5	210	12	0.05
	GS2	-6.5	-0.8	220	14	0.00
	GS3	-7.7	-0.9	256	15	0.00
	GS4	-6.7	-0.7	241	17	0.00
	GS5	-3.2	-0.5	206	17	0.03
moy		*-5.8*	*-0.68*	*227*	*15*	
Y						
	GS1	-11.9	-0.7	303	12	0.002
	GS2	-10.4	-0.7	253	14	0.001
	GS3	-12.6	-0.8	291	14	0.002
	GS4	-7.8	-0.7	231	16	0.002
	GS5	-15.4	-0.8	391	17	0.00
moy		*-11.6*	*-0.74*	*294*	*15*	

Tableau 2. Indices de régression linéaire (y=ax+b) calculés sur cinq groupes de souffle chez les locutrices D et Y. (a=coefficient de régression, r=coefficient de corrélation entre fréquences et syllabes; b=interception, dl=degré de liberté; p=pourcentage de probabilité pour que l'effet soit dû au hasard).

6. PROEMINENCES SYLLABIQUES DANS LES GROUPES RYTHMIQUES

Parmi les cinq groupes de souffle des deux locutrices, nous avons sélectionné deux groupes ayant la même fonction énonciative (adresse directe à l'interlocuteur par la même marque discursive *remarquez*), et contenant le même nombre virtuel de groupes rythmiques.

La transcription phonétique des deux groupes (G1 et Y1) a servi de base pour la segmentation en syllabes. (La transcription orthographique privée de signes de ponctuation permet une meilleure lisibilité.) L'étiquetage est fonction des propriétés morphologiques des éléments, correspondant grosso-modo à une distinction entre mots 'pleins' et mots 'grammaticaux'. L'étiquetage des syllabes en syllabes accentuables ('+') et non-accentuables ('.') a abouti à 7 'groupes rythmiques' ou 'groupes accentuels' virtuels chez les deux locutrices (voir entre autres [Mertens, 1993]). La structuration mélodique virtuelle des groupes rythmiques est visualisée par la typographie:

D1

vous avez remarqué qu'les parents disent toujours plus tard tu verras

[za e ʀmaʀ ke kle pa ʀɑ̃ di:z tu ʒu:ʀ ply ta:ʀ ty vɛ ʀɑ]

. + . + . . + + . + . + . . +

----/--------/----------/---/-------/---------/-------/

Y1

remarquez qu'il s'agit d'allure d'un homme très jeune il a vingt-cinq ans

[ʀəmaʀ ke kil sa ʒi dal ly:ʀ dɛ̃ ɔm tʀe ʒœn i la vɛ̃tsɛ̃kɑ̃]

. . + . . + . + . + . + . + . + . . +

--------/--------/------/--------/---------/-----/--------/

En comparant les deux groupes de souffle du point de vue des règles de la syntaxe de l'écrit, nous constatons qu'ils englobent, chez les deux locutrices, des structures syntaxiques différentes. Le découpage du corpus basé sur la syntaxe de l'écrit n'aurait certainement pas permis de les considérer comme des unités comparables.

Les figures 5 et 6 illustrent la réalisation (D1', Y1') des groupes rythmiques virtuels. On note que la subdivision de D1' en groupes rythmiques diffère considérablement de celle de Y1'. Y tend à marquer une montée de F0 sur la totalité des syllabes accentuables (Y1'), alors que D choisit une réalisation différente (D1'). Elle réalise la montée virtuelle sur la première syllabe accentuable [ke], ensuite elle superpose un contour plat sur les trois syllabes accentuables qui suivent [ʀɑ̃ di:z ʒu:ʀ] qu'elle sépare des deux dernières syllabes accentuables [ta:ʀ ʀɑ] par un décroché mélodique. Parmi les 7 proéminences virtuelles dans chaque groupe (D1 et Y1), 3 sont réalisées chez D et 6 chez Y:

D1'

[za e ʀmaʀ ke kle pa ʀɑ̃ di:z tu ʒu:ʀ ply ta:ʀ ty vɛ ʀɑ]

----------/----------/----------/

Y1'

[ʀəmaʀ ke kil sa ʒi dal ly:ʀ dɛ̃ ɔm tʀe ʒœn ila vɛ̃tsɛ̃kɑ̃]

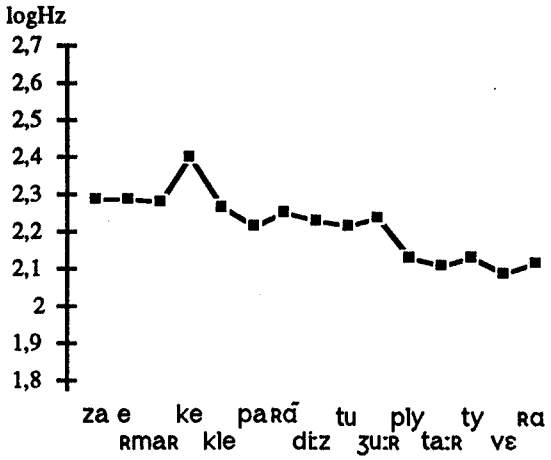
---------/---------/------/--------/---------/----------/

En résumé, on peut dire qu'à nombre de syllabes accentuables égales, la locutrice D réalise moins de proéminences dans un groupe de souffle que la locutrice Y. Cette tendance expliquerait pourquoi la pente des droites de régression, caractérisant la 'déclinaison' de F0 est moins 'abrupte' chez D. Si cet effet pouvait être démontré pour l'ensemble du corpus, on pourrait dire que la locutrice D privilégie une structuration globale en 'grands constituants', alors que la locutrice Y semblerait suivre une structuration en groupes rythmiques (discussion personnelle avec Vaissière et Kassai, 1993).

CONCLUSION

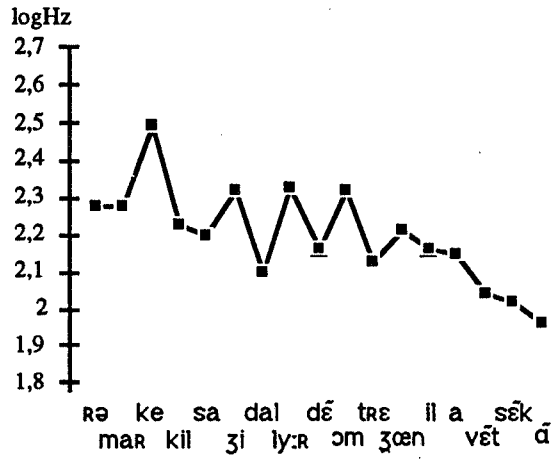
Nous avons tenté de montrer que les variations globales de la fréquence fondamentale, mesurées dans un extrait d'interview télévisée de Marguerite Duras et de Marguerite Yourcenar peuvent être comparées en terme

de 'déclinaison' de F0. Il s'est avérée que les différences entre les deux locutrices, concernant la gestion de la 'déclinaison' sont dûes à des différences de structuration en groupes rythmiques. L'existence du phénomène de 'déclinaison' dans ce corpus de parole non-lue a été démontrée grâce à la segmentation du corpus en unités discursives. Ce type de segmentation paraît plus approprié pour l'analyse de la parole non-lue que le découpage en phrases "grammaticales", car dans une argumentation élaborée (lue ou non-lue), chacun des énoncés possède son propre contour de 'déclinaison', (dû à son rôle dans l'argumentation), ce qui ne peut être ignoré lors de leur segmentation en unités.



(D1') vous avez remarqué que les parents disent toujours plus tard tu verras

Figure 5. Valeurs maximales de F0 par syllabe dans le groupe initial de paragraphe D1' (La typographie de la transcription phonétique ne suit pas la structuration mélodique.)



(Y1') remarquez qu'il s'agit d'allure d'un homme très jeune il a vingt-cinq ans

Figure 6. Valeurs maximales de F0 par syllabe dans le groupe initial de paragraphe Y1' (La typographie de la transcription phonétique ne suit pas la structuration mélodique.)

Je tiens à remercier Mme Jacqueline Vaissière, Mme Ilona Kassai et M. Ivan Fonagy pour leurs commentaires apportés à une version précédente de cet article.

BIBLIOGRAPHIE

BROWN, G. - CURRIE, K.L. - KENWORTHY, J. (1980), Questions of intonation, Croom Helm, London.

BRUCE, G. (1982), Textual Aspects of Prosody in Swedish, Phonetica, 39, 274-287.

DEESE, J. (1980), Pauses, prosody, and the demandes of production in language, in: Dechert H.W. - Raupach, M. (eds), TEMPORAL VARIABLES IN SPEECH: STUDIES IN HONNOR OF FRIEDA GOLDMAN-EISLER, Mouton, The Hague.

DUEZ, D. (1991), La pause dans la parole de l'homme politique, Editions du CNRS, Marseille.

FAGYAL, Zs. (1994), Le style vocal de Marguerite Duras, in: MARGUERITE DURAS: RENCONTRES DE CERISY, A. Vircondelet (ed.), Paris, Ecriture, 69-82.

GELFER, C. G. - HARRIS, K. S. - COLLIER, R. - BAER, T. (1983), Speculations on the control of fundamental frequency declinaison, Status Report on Speech Research, 76, Haskins Laboratories, 51-63.

GOLDMAN-EISLER, F. (1968), Psycholinguistics: Experiments in Spontaneous Speech, Academic Press, London and New York.

HAUSER, M. D. - F0WLER, A. (1992), Fundamental frequency declinaison is not unique to human speech: Evidence from human primates, JASA, 91 (1), January, 363-369.

KREIMAN, J. (1982), Perception of sentence and paragraph boundaries in natural conversation, Journal of Phonetics, 10, 163-175.

LEHISTE, I. (1975), The Phonetic Structure of Paragraphs, in: Structure and Process in: Cohen, A. - Noteboom, S. (eds.), SPEECH PERCEPTION, Springer Verlag, 195-206.

LIEBERMAN, P. (1967), Intonation, Perception and Language, MIT Res. Monogr., 38, MIT Press, Cambridge.

LIEBERMAN, P. et al. (1985), Measures of sentence intonation of read and spontaneous speech in American English, JASA, 77 (2), February, 649-657.

MAEDA, S. (1976), A Characterization of American English Intonation, Doctoral Dissertation, MIT, Cambridge.

PIERREHUMBERT, J. (1979), The perception of fundamental frequency declinaison, JASA, 66 (2), August, 363-369.

SWERTS, M. (1991), An experimental approach to the study of prosody in spontaneous speech, Proceedings of the ETRW 'Phonetics and Phonology of Speaking Styles', Barcelona, 511-515.

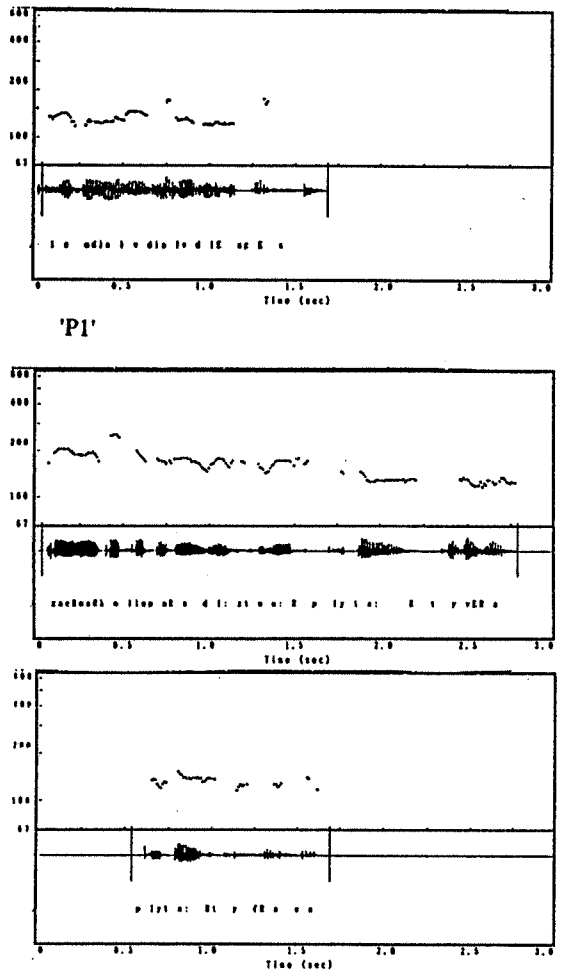
THORSEN, N. G. (1985), Intonation and text in Standard Danish, JASA, 77 (3), March, 1205-1215.

UMEDA, N. (1982), "F0 declination" is situation dependent, Journal of Phonetics, 10, 279-290.

VAISSIERE, J. (1983), Language-Independent Prosodic Features, in: Cutler, A. - Ladd, R. D. (eds.), PROSODY: MODELS AND MEASUREMENTS, Springer Verlag, 53-66.

ANNEXE

LOCUTRICE D:



'P1'

'P2'

Figure 7. Extrait du corpus D. (Fin du paragraphe 1 'P1'): le jeu d'la loi d'la loi d'l'espèce # (Paragraphe 2 'P2'): 's a'ez r'marqué qu'les parents disent toujours plus tard tu verras plus tard tu f'ras ça [...] (#=pauses)

LISTE DES GROUPES DE SOUFFLE FIGURE 5:

GS1: et cette c'est cette séparation du corps du jeune homme mort

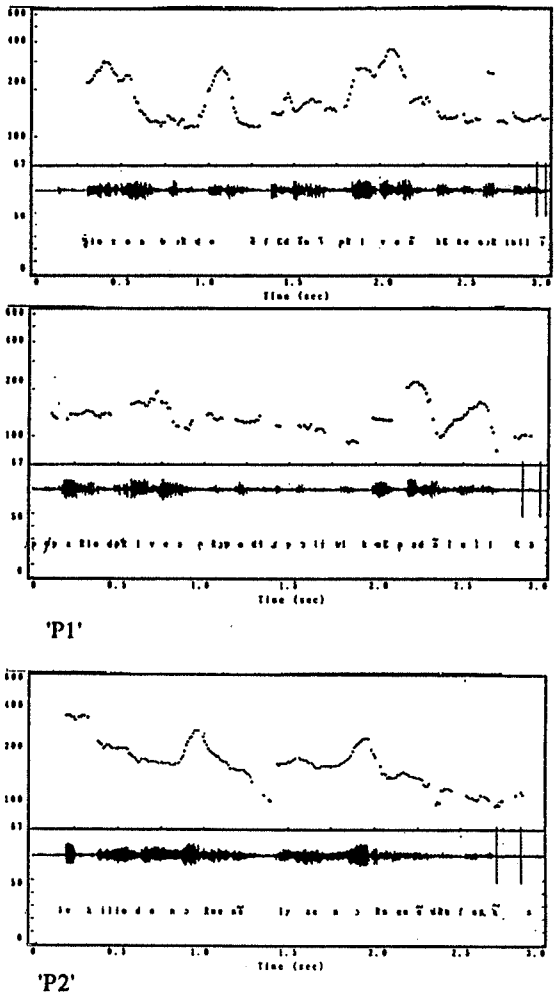
GS2: vous avez r'marqué qu'les parents disent toujours plus tard tu verras

GS3: si vous vous souv'nez dans le livre je l'dis à un moment donné

GS4: 'savez quand les choses ne sont pas dites 'les n'ont pas été dites là de ma part

GS5: elle a pas quitté elle a pas quitté elle a continué à en(seigner oui) [chevauchement]

LOCUTRICE Y:



'P1'

'P2'

Figure 8. Extrait du corpus Y. (Paragraphe 1 'P1'): j'les ai abordés m'enfin dans mon privé si l'on peut parler d'privé à propos d'la politique mais pas dans les livres # (Paragraphe 2 'P2'): quoi qu'il y ait énormément i' a énormément d'références # [...] (#=pause)

LISTE DES GROUPES DE SOUFFLE FIGURE 4:

GS1: et peut-être plus encore avec le mensonge politique

GS2: mais tout d' même le seul homme qui ait pris l'deuil pour une impératrice

GS3: quoi qu'il y ait énorme il y a énormément d'références

GS4: remarquez qu'il s'agit d'allure d'un homme très jeune il a vingt-cinq ans

GS5: bien plus j'avais même l'impression qu' c'était arrivé que ça s' passerait

SEGMENTATION LEXICALE EN FRANCAIS PARLE
La liaison enchaînée.

S. WAUQUIER-GRAVELINES

Laboratoire de Psychologie Expérimentale,
Equipe de Psycholinguistique
28 rue Serpente
75005 PARIS.

U.F.R de Linguistique Théorique et formelle
Université Paris 7
2, place Jussieu
75005 PARIS

Résumé.
Les principaux problèmes théoriques posés par la segmentation de la chaîne parlée sont rapidement évoqués ici. Cette question est ensuite abordée plus spécifiquement grâce à l'étude d'un phénomène très répandu en français : la liaison enchaînée, qui permet d'interroger de manière très intéressante les modèles d'accès au lexique actuellement proposés. Les résultats obtenus dans deux expériences sur la détection en temps réel de la consonne de liaison nous invitent à remettre en cause les modèles de segmentation pré-lexicaux actuels pour lesquels l'information segmentale seule est utilisée lors de l'accès au lexique.

Mots-clés psycholinguistique, segmentation, accès au lexique, phonologie, parole continue.

1 SEGMENTATION ET ACCES AU LEXIQUE :

Pour "comprendre" un discours oral, un auditeur doit isoler et reconnaître les mots qui le constituent. Or les caractéristiques du signal de parole présentent un défi permanent à cette opération. La parole est variable : chaque son produit a une réalisation phonétique propre et non exactement reproductible, dont les paramètres sont nombreux et eux-mêmes variables selon les locuteurs. La parole est continue : contrairement à un texte écrit structuré par des blancs et alinéas qui isolent les mots, un signal de parole ne comporte pas de marque universelle, systématique et récurrente permettant de repérer les frontières de mots, de morphèmes ou de segments. Si l'on peut repérer sur un spectrogramme des indices de frontières, ces indices varient d'une langue à l'autre (Vaissière, 1983). Face à ce flux d'informations séquentiel et continu, comment procède l'auditeur pour dégager dans le continuum acoustique des repères correspondant aux représentations lexicales? Les recherches sur ce thème s'articulent sur deux axes :
-Quelles sont les procédures de segmentation utilisées par les auditeurs ?
-Existe-t-il des unités infra-lexicales facilitant cette opération de segmentation et dans ce cas quelles sont-elles ?

Ces problèmes sont évidemment intriqués : la procédure de segmentation qu'adopte un auditeur est générée par les représentations mentales qui sont à sa disposition et vice-versa

Certaines conceptions théoriques dites post-lexicales, considèrent qu'il n'existe pas de procédure de segmentation explicite lors de l'accès au lexique (Bard, 1991, McClelland et Ellman, 1986). Dans ces modèles, les unités lexicales sont activées par l'information contenue dans le signal. Le système élabore des hypothèses à partir de cette information ; la dynamique d'inhibition et d'activation des candidats à partir de cette information, fait émerger le mot concerné et élimine les autres concurrents. Ceci suppose donc que l'auditeur ne cherche pas à élaborer un format correspondant au mot à reconnaître *avant* de l'avoir reconnu. La segmentation n'est qu'un résidu de l'accès au lexique. En d'autres termes, la frontière qui sépare un mot d'un autre apparaît d'elle même quand le mot est reconnu.

Pour un autre courant de recherche au contraire, il est nécessaire de segmenter le signal afin de permettre l'accès au lexique. Autrement dit, la segmentation est un processus pré-lexical : il faut établir des frontières entre les mots pour pouvoir les reconnaître. Dans cette perspective se pose alors le problème de la nature et de la taille de l'unité de segmentation.

Personne ne soutient plus actuellement l'hypothèse selon laquelle le phonème serait une unité de segmentation de la parole continue en temps réel. Phonétiquement d'une part les phonèmes sont la plupart du temps coarticulés et donc difficilement isolables. Leurs réalisations sont conditionnées par l'environnement et donc très variables.
La syllabe présente des caractéristiques articulatoires (articulations consonnantiques autour d'un noyau vocalique correspondant aux ouvertures et fermetures du canal buccal réduisant la coarticulation entre les syllabes) et une structure phonologique qui en font un candidat plus sérieux. Des recherches nombreuses laissent penser que la syllabe a une réalité psychologique : elle est reconnue très précocement chez les nouveaux-nés (Bertoncini & Mehler, 1981) et beaucoup plus spontanément utilisée que le phonème

chez des illetrés pour scander du matériel acoustique (Morais, Content, Cary, Mehler & Segui, 1989)
Pour les langues romanes et qui plus est les langues romanes sans accent lexical distinctif comme le français, des résultats expérimentaux laissent penser que les auditeurs utilisent une segmentation en syllabes pour accéder au lexique (Mehler, Dommergues, Frauenfelder & Segui, 1981), il semblerait que ce soit également le cas pour l'espagnol et le portugais. Ces résultats ne sont cependant pas généralisables à toutes les langues et posent problème pour des langues à accent comme l'anglais par exemple (Cutler, Mehler, Norris & Segui, 1986) Les auditeurs de cette langue semblent plutôt s'appuyer sur la succession de syllabes accentuées et inaccentuées pour découper le signal en utilisant les syllabes fortes comme débuts potentiels de mots (Cutler & Norris, 1988). Par ailleurs des recherches ont montré (Otake, Giyoo, Cutler & Mehler, 1993) que les japonais utilisaient plus volontiers la more que la syllabe pour segmenter leur langue. Ces résultats ont suggéré qu'il pourrait exister une stratégie métrique de segmentation propre à chaque langue : Le français est traditionellement considéré comme une langue "syllable-timed" où l'unité rythmique de base de la langue serait la syllabe. Quelles que soient les choix théoriques et empiriques que l'on adopte pour définir précisèment l'isosyllabicité (Bertinetto, 1989) et malgré les nombreux problèmes que pose cette classification des langues en "stress-timed" et "syllabled-timed" on peut considérer que la syllabe en français est effectivement une unité saillante de structuration rythmique du continuum sonore notamment par rapport à des langues comme l'anglais où la régulation rythmique de la parole s'effectue par la succession de syllabes accentuées et non-accentuées. Considérant ces caractéristiques rythmiques du français, l'hypothèse de segmentation syllabique s'avère dans un premier temps tout à fait défendable. Pourtant, dès que l'on examine de la parole continue, cette conception de la segmentation prélexicalè en syllabes se heurte aux phénomènes phonologiques de sandhis tels que la liaison ou l'enchaînement qui créent des suites à segmentation ambigüe (ex petit ami / petit tamis).

Ces problèmes sont directement liés au fait que les hypothèses de segmentation ont toutes tendance à considérer que les mots, dans la parole continue, sont ce que Grosjean et Gee (1987) appellent "WD Words" (Written dictionnary words), des représentations lexicales abstraites et identiques à elle-mêmes quel que soit le contexte phonologique dans lequel elles apparaissent. Une autre hypothèse de segmentation a été récemment proposée pour le français (Christophe, 1993): celle d'une segmentation prélexicale qui définirait des domaines prosodiques tels que le groupe clitique (tête lexicale + clitique) ou la phrase phonologique (tête lexicale + côté non récursif soit la gauche en français) à l'intérieur desquels aurait lieu l'accès au lexique. Ces unités définies par la phonologie prosodique (Nespor &Vogel, 1986) à partir de la théorie syntaxique X, présente en effet le double avantage de prendre en considération les règles d'organisation prosodiques, phonologiques et syntaxiques du continuum sonore et donc de ne pas imposer d'emblée une segmentation par les "mots". Comme le soulignent Grosjean et Gee (1987) : " The WD word (as a linguistic unit in its own right, as an ensemble of smaller units, or as an acoustic-phonetic segment)may not after all be the ideal domain over which lexical processing takes place. S'il est évident que la segmentation de la parole a pour but de retrouver des unités lexicales abstraites, rien n'indique que le traitement du signal se fasse par des unités abstraites inférieures ou égales au mot et non à partir de portions de signal supérieures au mot et définies par des critères autres que des critères orthographiques. Des phénomènes phonologiques tels que la liaison enchaînée en français posent spécifiquement ce problème.

1-2 La liaison enchaînée en français ambiguité de frontières et segmentation

Le français a tendance à réaliser des syllabes rythmiquement stables où la réduction vocalique se produit dans une proportion bien moindre qu'en anglais. Par ailleurs le français est une langue où la proportion de syllabes ouvertes de type C((C)C)V est très élevée (Léon, 1964). Cette préférence pour les syllabes ouvertes entraîne le plus souvent pour des syllabes fermées suivies d'une syllabe à attaque vide (ex une orange) un enchaînement, c'est à dire un passage de la coda consonnantique à l'attaque vide de la syllabe suivante. Le français par ailleurs est une langue qui tolère mal les hiatus. La liaison enchaînée est un cas particulier d'enchaînement où une consonne dite flottante se réalise à l'attaque de la syllabe suivante mais uniquement dans ces conditions (ex [poeti] (petit) et [an] (âne) quand les mots sont prononcés isolèment mais [poetitan] (petit âne) quand ils sont prononcés à la suite l'un de l'autre). Cette apparition d'un segment flottant réduit le hiatus. La resyllabation de la consonne de liaison se fait sur une frontière de mots et la consonne qui se réalise là a un statut unique : elle appartient lexicalement au premier mot dans le sens où sa nature phonétique est déterminée par lui et varie avec lui (petit/petite), mais ne se réalise que sur la structure syllabique du mot suivant. La proportion des liaisons enchaînées en français est extrêmement élevée (Encrevé, 1988).
On classe traditionnellement les liaisons en "facultatives" et "obligatoires" (Nyrop, 1902; Fouché 1959; Delattre, 1966; Encrevé, 1988; De Jong, 1991). La fréquence et la distribution des liaisons est contrainte par la taille et la nature du domaine phonologique dans lequel elle se produit (Nespor&Vogel, 1986; Encrevé, 1988; De Jong, 1991). Ainsi la liaison est considérée comme obligatoire dans les groupes clitiques et les phrases phonologiques et réalisée respectivement à 99% et à 75% par les locuteurs (De Jong,1991). On constate donc que le degré de contrainte qui conditionne la

réalisation des liaisons dans un domaine est proportionnelle au degré de cohésion de ce domaine.

En faisant apparaître un segment resyllabifié sur la frontière entre deux mots à l'intérieur d'un domaine phonologique, la liaison enchaînée crée dans ce domaine une syllabe dont les deux constituants n'appartiennent pas à la même unité lexicale et donc gomme les limites de mots. On peut faire l'hypothèse que ce phénomène pose un problème crucial à l'auditeur : quelles stratégies adopte-t-il pour segmenter, quand le mot auquel il veut accéder est réalisé avec à l'attaque de la syllabe initiale un segment qui n'appartient pas à sa structure sous-jacente. Ce segment resyllabifié trace la frontière potentielle d'un mot qui n'existe pas dans la plupart des cas (le mot "tane" dégagé de "petit âne" n'existe pas en français). Ceci pose évidemment un problème de segmentation : dans les deux experiences qui suivent, nous avons tenté d'observer en temps réel comment les sujets traitent ce segment qui crée une syllabe sur la frontière entre deux mots.

2 EXPERIMENTATION

2.1 EXPERIENCE 1

Comme nous l'avons vu, la liaison enchaînée fait apparaître un segment sur la frontière entre deux mots et la rend ambigüe : dans une suite comme petit éléphant, l'auditeur doit repérer la nature spéciqfique de ce segment afin de ne pas segmenter de manière erronée (petite éléphant ou petit téléphant). Dans cette expérience nous avons utilisé la procédure de détection de phonème généralisée (Dphg) proposée par Frauenfelder et Segui (1989), qui consiste à faire détecter des phonèmes pouvant apparaître dans n'importe quelle position à l'intérieur d'une séquence afin d'observer si la détection d'un phonème cible "flottant" généré par une liaison diffère de celle d'un phonème cible appartenant à la représentation canonique d'un mot.

L'hypothèse était la suivante : si les sujets manifestent une plus grande difficulté à détecter ce phonème, ceci suppose que ce segment impose un traitement spécifique, qu'il doit être à la fois identifié comme un phonème de liaison et différencié d'un phonème initial de mots. La frontière de mot doit être calculée en tant que telle.

Matériel

8 paires d'items ont été construites faisant respectivement apparaître un /t/ initial et un /t/ de liaison resyllabifié (type grand éléphant/ grand téléphone ou élégant iranien / élégant tirailleur) dans des phrases phonologiques. Les facteurs suivants ont été contrôlés.

1) la fréquence : les mots constituant les paires tests grand téléphone / éléphant étaient de fréquence sensiblement identiques. Les fréquences ont été mesurées à partir des tables du programme Brulex.

2) les mots ont en commun la structure syllabique et la nature des segments dans au moins les deux premières syllabes afin de tenter d'éviter que la coarticulation du /t/ et de la voyelle subséquente entraîne une réalisation de la consonne très différente dans chacun des items de la paire.

Ces phrases tests ont été insérées dans deux listes contrebalancées de 70 phrases de remplissage comportant soit le son /t/ à l'initiale des mots (tapis), à l'intérieur des mots (beauté, attaché), à la fin des mots (Colette), soit aucun son /t/. Le matériel est donc constitué de deux blocs de 78 phrases. Les phrases ont été lues par un locuteur masculin de langue maternelle française. Elles ont été enregistrées sur un magnétophone à bandes REVOX 516, puis transférées par digitalisation sur le disque dur d'un Macintosh SII où elles ont été traitées par le programme Audiomédia.

Les sujets sont 23 étudiants en licence de Psychologie à l'Université René Descartes, Paris V. La tâche était la suivante : appuyez sur le bouton réponse dès que vous entendez le son /t/ quel que soit son emplacement dans la phrase. Les sujets suivaient avant l'expérience une session d'entraînement à l'issue de laquelle il leur était explicitement demandé de détecter aussi les liaisons quand ils ne l'avaient pas fait dans les phrases d'entraînement. L'expérience elle-même démarrait ensuite.

Résultats

T liaison	T initial	différence
619 ms	573 ms	46 ms

Tableau 1 : Temps de détection des consonnes initiales et liaisons.

On constate (tableau 1) une différence significative de temps de réaction tant par items ($t(1, 8) = 2.87$ $p=0.02$) que par sujets ($t(1,19)= 2.89$ $p= 0.009$) . Tous les temps de réactions inférieurs à 100 ms ainsi que tous ceux supérieurs à 2000 ms ont été éliminés. Enfin on constate 15 %de t de liaisons non-détectées pour 5% de t initiaux non-détectés.

Ces résultats confirment l'hypothèse que nous avions faite selon laquelle les segments générés par la liaison imposent un traitement spécifique nécessaire à la définition des frontières de mots.

Comme nous l'avons indiqué, la distribution des liaisons est liée à la taille et à la nature du domaine dans lequel elle apparaît. Comme le signale De Jong (1991), les contraintes d'apparition de la liaison dans le groupe clitique sont plus fortes que dans la phrase phonologique : dans le groupe clitique, la liaison est produite 99 fois sur 100. Cette contrainte s'applique-t-elle également en perception ?

Dans une deuxième expérience, nous avons tenté de voir si l'on observe également un traitement spécifique de la consonne de liaison et la nécessité

d'un calcul des frontières de mots dans des groupes clitiques

2.2 EXPERIENCE 2

Matériel

8 paires d'items ont été construites faisant respectivement apparaître un /n/ initial et un /n/ de liaison resyllabifié (type un navire / un avion ou un oiseau / un noyau) dans des groupes clitiques. Les facteurs suivants ont été contrôlés.
1) la fréquence : les mots constituant les paires tests navire / avion étaient de fréquence sensiblement identiques. Les fréquences ont été mesurées à partir des tables du programme Brulex.
2) les mots ont en commun autant que possible la structure syllabique et la nature des segments dans au moins la première syllabe afin de tenter d'éviter que la coarticulation du /n/ et de la voyelle subséquente entraîne une réalisation de la consonne très différente dans chacun des items de la paire .
Ces phrases tests ont été insérées dans deux listes contrebalancées de 70 phrases de remplissage comportant soit le son /n/ à l'initiale des mots (noisette), à l'intérieur des mots (animal), à la fin des mots (benne), soit aucun son /n/. Le matériel est donc constitué de deux blocs de 78 phrases. Les phrases ont été lues par un locuteur masculin de langue maternelle française. Elles ont été enregistrées sur un magnétophone à bandes REVOX 516, puis transférées par digitalisation sur le disque dur d'un Macintosh SII où elles ont été traitées par le programme Audiomédia.

Les sujets sont 42 étudiants en licence de Psychologie à l'Université René Descartes, Paris V. La tâche était La suivante: appuyez sur le bouton réponse dès que vous entendez le son /n/ quel que soit son emplacement dans la phrase. Les sujets suivaient avant l'expérience une session d'entraînement à l'issue de laquelle il leur était explicitement demandé de détecter aussi les liaisons quand ils ne l'avaient pas fait dans les phrases d'entraînement. L'expérience elle-même démarrait ensuite.

Résultats

Les sujets ont manifesté tout au long de la passation, une grande difficulté à exécuter la tâche. Malgré l'augmentation du nombre de sujets, trop peu d'items de liaisons ont détectés pour que la différence en temps de réaction soit informative. Les erreurs de détection ont alors été observées systématiquement

N liaison	N initial
40% de non-détection	14% de non détection

Tableau 2 : Pourcentages de non-détection des consonnes initiales et de liaison.

On constate (tableau 2) une différence significative du pourcentage d'items détectés dans les deux conditions ($t(1, 41) = 5.75$ $p=0.00002$)
Bien que la consigne après la session d'entraînement ait demandé explicitement aux sujets qui ne l'avaient pas fait de détecter aussi les liaisons et confirmé qu'il fallait bien le faire à ceux qui l'avaient fait, on constate un grand nombre d'absence de détection de la consonne de liaison dans les groupes clitiques.

On peut remarquer que la non-détection du /n/ est également importante quand il s'agit du /n/ initial (14%). On remarque également que pour le /n/ dans les deux conditions le nombre de non-détection est respectivement plus important que pour le /t/ dans l'expérience 1 (15% pour le /t/ de liaison et 5% pour le /t/ initial). Ces résultats sont sans doute en partie attribuables à la nature acoustique du phonème à détecter. L'occlusive alvéolaire /t/ présente des caractéristiques acoustiques qui la rendent plus saillante que le /n/. Mais si le /n/ apparaît globalement plus difficile à détecter, ceci n'explique pas la différence entre les deux conditions de l'expérience 2. Les sujets ont manifesté une grande difficulté à reconnaître et isoler les phonèmes de liaison en groupes clitiques

Discussion

Ces résultats confirment que les sujets ont eu plus de difficultés à détecter la consonne de liaison que la consonne initiale de mot aussi bien en groupe clitique qu'en phrase phonologique. On peut proposer deux explications non exclusives l'une de l'autre permettant d'interpréter ces résultats.
a-Les deux réalisations acoustiques de la consonne, en contexte de liaison et en début de mot sont différentes et les sujets manifestent par leur temps de réaction leur sensibilité à cette différence. On peut alors considérer que lors de la segmentation de la chaîne parlée, les sujets utilisent cette information acoustique pour établir les frontières de mots et frontières de morphèmes. L'idée qu'il pourrait s'agir uniquement d'un phénomène de bas niveau est assez peu recevable : la variabilité du signal de parole et le peu de certitudes qu'apportent les études sur les invariants phonétiques peuvent effectivement laisser penser que l'information acoustico-phonétique n'apporte pas une information suffisante pour désambiguiser une frontière de mot. Enfin , toute la littérature phonétique et phonologique (Encrevé,1988) concernant la liaison enchaînée considère qu'il n'existe pas de différences de réalisation entre ces deux types de segments. Une analyse spectrographique des phrases est actuellement en cours.
b-La différence de temps de réaction peut également être attribuée au fait que les sujets différencient plutôt deux représentations sous-jacentes. Ces deux segments ont effectivement un statut phonologique différent : Le /t/ ou le /n/ initial sont des segments appartenant à la représentation lexicale sous-jacente et qui se réalisent quel que soit le contexte dans lequel apparaît le mot alors que la réalisation de la consonne

de liaison est entièrement régie par des règles phonologiques spécifiques du français. De là découle un problème de segmentation. La présence d'un phonème de liaison génère en effet une séquence qui ne correspond la plupart du temps à aucun mot de la langue. La frontière syllabique crée la possibilité d'un début de mot qui n'existe pas ("tarbre" dans grand arbre). Le sujet ne peut pas retrouver le mot sans en recalculer la frontière en clivant la syllabe. Les hésitations et erreurs de production que l'on constate chez les enfants de deux à trois ans ("le nane" pour l'âne) confirmeraient qu'un locuteur du français doit nécessairement apprendre explicitement à repérer les frontières de mots et les contraintes qui régissent l'apparition de la liaison. Ceci n'exclut pas pour autant que l'information acoustico-phonétique joue un rôle. Les stratégies de segmentation employées par les auditeurs dériveraient de leur connaissance des règles de syllabification de la langue concernée, elles mêmes élaborées à partir d'une interaction étroite entre les informations acoustico-phonétiques et la connaissance phonologique qu'un auditeur a de sa langue. Ces deux types d'informations seraient confrontées à chaque instant pour structurer et segmenter le signal.

Enfin, les règles de syllabification en français (liaisons, élisions, réduction du schwa) ne s'appliquent pas de manière aléatoire sur le continuum sonore mais sont tributaires des domaines phonologiques spécifiques (principalement le groupe clitique et la phrase phonologique) auxquels elles s'appliquent. La phonologie prosodique (Nespor &Vogel, 1986) pose que le degré de contrainte qui conditionne l'application d'une règle phonologique dans un domaine est proportionnelle au degré de cohésion de ce domaine. Bien que les résultats de l'expérience 1 et de l'expérience 2 ne soient pas directement comparables, il apparaît que les sujets manifestent une difficulté à détecter les consonnes de liaisons et ce spécifiquement dans les groupes clitiques. Ceci suggère que plus la cohésion du domaine d'où le sujet doit extraire un son est grande plus la tâche est difficile. Ces résultats montrent comment les hypothèses de stratégie de segmentation lexicale et post-lexicale proposées actuellement en psycholinguistique achoppent sur les phénomènes de "sandhis" dont la liaison constitue un exemple. Ceci laisse également penser , comme l'a suggéré A. Christophe (1993) que la segmentation de la parole continue pourrait s'effectuer à partir de fragments de signal plus important que le mot comme les domaines phonologiques.à l'intérieur desquels s'effectuerait l'acès lexical. Cette hypothèse est rarement envisagée en perception de la parole alors que de nombreux faits empiriques dans des langues présentant des caractéristiques phonologiques très différentes aussi bien que certaines théories linguistiques apportent des arguments qui la rendent très plausible.

BIBLIOGRAPHIE

BARD, E.G.(1991) Competition, Lateral Inhibition and Frequency : Comments on the Chapters of Frauenfelder and Peeters, Marslen-Wilson and others In G.T. M Altman (ed.) *Cognitive models of Speech Processing: Psycholinguistic and Computational Perspectives*, pp 263-280, Cambridge, Ma : The MIT Press.

BERTINETTO, P.M. (1989) . Reflections on the dichotomy "stress"vs."syllable timing"*Revue de Phonétique Appliquée , 91-92-93.*

BERTONCINI, J & MEHLER, J.(1981). Syllables as units in infant speech perception, *Infant Behavior and Development,4,* pp 247-260.

CHRISTOPHE, A. (1993). *Rôle de la prosodie dans la segmentation en mots,* Thèse de Doctorat, EHESS Paris.

CUTLER , A. & NORRIS , D. (1988). The Role of strong syllables in segmentation for lexical access. *Journal of Experimental Psuychology , 14,* pp 113-131.

CUTLER, A. , MEHLER , J., NORRIS , D.& SEGUI ,J. (1986). The Syllable's Differing Role in the Segmentation of French and English . *Journal of Memory and Language, 25,* pp 385-400.

DE JONG, D. (1991). La liaison à Orléans (France) et à Montréal (Québec). In *Actes du XIIème Congrès international des Sciences Phonétiques*, Août 1991, Aix-en-Provence , pp 198-201.

DELATTRE, P., (1966). *Studies in French and Comparative Phonetics,* La Haye, Mouton.

ENCREVE, P. (1988) *La liaison avec et sans enchaînement,* Seuil, Paris.

FRAUENFELDER, U. & SEGUI, J. (1989). Evidence for associative context effects in phoneme monitoring, *Memory and Cognition, 17,* pp134-140.

GROSJEAN, F & GEE, J.P. (1987). Prosodic Structure and Spoken Word Recognition. In U.H.Frauenfelder & L Komisarjewsky-Tyler (eds), *Spoken Word Recognition Cognition, Special Issue, Vol 25,* pp. 135-157.

FOUCHE, P., (1959). Traité de prononciation française, Paris, Klinsieck, 2ème éd.

LEON, P.& M (1964) Introduction à la Phonétique corrective, Paris, Hachette-Larousse.

Mc CLELLAND, J.L. & ELMAN J.L. (1986) Interactive Processes in Speech Perception: The TRACE Model. In J.L. Mc Clelland, D.E. Rumelhart, & the PDP Research Group (Eds), *Parallel Distributed Processing, Vol 2: Psychological and Biological Models*, pp 58-121. Cambridge, MA: The MIT Press.

MEHLER, J., DOMMERGUES, J.Y., FRAUENFELDER, U. & SEGUI, J. (1981) . The syllable's Role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior, 20,* pp 298-305.

MORAIS, J. , CONTENT, A. , CARY, L. MEHLER, J.& SEGUI, J.(1989). Syllabic Segmentation and litteracy, *Language and Cognitive Processes ,4,* 57-67.

NESPOR, M. & VOGEL, I. (1986). *Prosodic Phonology,* Dordrecht: Foris.

NYROP, K., (1902). Manuel phonétique du français parlé. (Traduction française, Paris, Picard)

OTAKE , T. GIYOO, H. CUTLER, A & MEHLER, J. (1993) . Mora or syllable ? Speech segmentation in japanese, *Journal of Memory and Language, vol 32,* 2.

VAISSIERE, J (1983) Language independant Prosodic Features in A. Cutler & R. D. Ladd (eds) *Prosody : Models and Measurements, Springer series in Language and Communication 14,* pp. 53-66. Berlin Springer Verlag.

LA CHAÎNE PHONÉMIQUE ET LE PATRON ACCENTUEL LEXICAL SONT-ILS REPRÉSENTÉS SÉPARÉMENT ? UNE ÉTUDE D'ILLUSIONS PERCEPTIVES

VALÉRIE PASDELOUP*, RÉGINE KOLINSKY ET JOSÉ MORAIS

LABORATOIRE DE PSYCHOLOGIE EXPÉRIMENTALE,
UNIVERSITÉ LIBRE DE BRUXELLES, 117 AV. A. BUYL, 1050 BRUXELLES
*ADRESSE PERMANENTE : 8 RUE DU CHÂTEAU-LANDON, 75010 PARIS

Résumé

L'objectif de cette étude est d'examiner si lors du traitement perceptif le patron accentuel et la chaîne phonémique sont représentés séparément dans une langue à accent lexical. Nous avons utilisé une tâche de détection de mots- et pseudomots-cibles en situation d'écoute dichotique : deux stimuli sont présentés simultanément, l'un à l'oreille droite et l'autre à l'oreille gauche. Nous avons testé si le sujet peut percevoir illusoirement la cible en combinant l'information phonémique du stimulus présenté à une oreille avec l'information du patron accentuel du stimulus présenté à l'autre oreille. Les résultats suggèrent que le sujet peut entendre une telle illusion, c'est-à-dire que le patron accentuel et la chaîne phonémique peuvent être représentés séparément. Cependant, l'apparition de ces illusions perceptives dépend du type de patron accentuel (par défaut vs. marqué) présenté.

Mots-clés : accent lexical, prosodie, traitement perceptif, représentations phonologiques.

1. INTRODUCTION

Dans les théories phonologiques classiques, ou linéaires, l'information relative à l'accent est spécifiée pour chaque syllabe du mot par le trait [±accent], alors que dans les théories phonologiques non-linéaires les éléments suprasegmentaux sont indépendants de la chaîne segmentale à laquelle ils sont associés. La phonologie métrique postule l'autonomie du patron accentuel lexical et de la chaîne phonémique à laquelle il est associé /1/, et la phonologie autosegmentale postule l'autonomie de la ligne tonale et de la ligne segmentale /2/. Cette autonomie est attestée dans les langues tonales, où l'on observe que plusieurs tons peuvent se réaliser sur une seule syllabe et, inversement, qu'un ton peut se diffuser sur plusieurs syllabes ; de même, si un mot perd une syllabe, le ton qui était porté par cette syllabe est associé à une autre syllabe. L'exemple le plus probant de cette autonomie s'observe dans des jeux de mots qui consistent à permuter les deux syllabes d'un mot ; dans ce cas, les tons ne sont pas affectés par les permutations segmentales : ils restent associés à leur position syllabique originelle /3/.

L'objectif du présent travail est de tenter de vérifier expérimentalement l'hypothèse de séparabilité perceptive de la chaîne phonémique et du patron accentuel lexical : ces informations constituent-elles des représentations séparées, ou au contraire restent-elles constamment associées ? Nous avons ainsi examiné si en portugais, une langue à accent lexical, le patron accentuel et la chaîne phonémique sont représentés séparément à un certain niveau du traitement.

Pour ce faire nous avons utilisé une technique qui a été établie à l'origine dans notre laboratoire pour l'étude du traitement de constituants phonologiques tels que les traits, segments, et syllabes /4, 5/. Cette technique utilise une tâche de détection de mots et de pseudomots-cibles en situation d'écoute dichotique. On présente simultanément deux stimuli auditifs (mots ou pseudomots), l'un à l'oreille droite et l'autre à l'oreille gauche. Les stimuli sont construits de telle manière que si le sujet combine une partie de l'information du stimulus présentée à une oreille avec une partie de l'information du stimulus présentée à l'autre oreille, il a l'illusion d'entendre un mot ou un pseudomot qui correspond à la cible, mais qui en fait ne lui a pas été présenté. L'apparition de cette illusion tend à prouver que les constituants présumés de chacun des stimuli sont représentés séparément à un moment donné du traitement perceptif, sinon ils ne pourraient pas se combiner entre eux. En effet, l'illusion requiert à la fois la séparabilité des constituants présumés et l'instabilité de ces constituants au regard de leur origine spatiale (oreille gauche ou droite). Cette technique s'inspire des travaux réalisés sur les erreurs de perception de parole appelées "phonetic feature blendings" : quand la syllabe /ba/ est présentée à une oreille et la syllabe /ta/ simultanément à l'autre oreille, le sujet entend souvent /da/ ou /pa/ /6/. Ces illusions perceptives résulteraient d'une erreur de combinaison entre le trait [±voisé] provenant du stimulus présenté à une oreille et le trait de lieu d'articulation provenant du stimulus présenté à l'autre oreille.

L'hypothèse selon laquelle le patron accentuel et la chaîne phonémique peuvent être représentés séparément

lors du traitement perceptif conduit à prédire que le sujet peut percevoir illusoirement un mot ou un pseudomot en combinant la chaîne phonémique du stimulus présenté à une oreille avec le patron accentuel du stimulus présenté à l'autre oreille. Dans une étude préalable en portugais, nous avons observé que le sujet peut entendre une telle illusion /7, 8/. Le patron accentuel et la chaîne phonémique semblent donc être représentés séparément à un moment donné du traitement perceptif. Cependant, l'apparition de ces erreurs de combinaison pourrait être soumise à des contraintes spécifiques d'association des constituants phonologiques. En effet, le phénomène illusoire n'avait lieu que dans le cas où les stimuli dichotiques étaient des mots partageant la 2ème syllabe.

Dans l'étude actuelle, nous avons repris la même situation expérimentale, mais nous avons essayé d'étudier de façon plus approfondie la nature de ces contraintes d'association. Nous avons donc testé si l'apparition des erreurs de combinaison dépend (i) du nombre et de la nature des segments phonémiques que partagent les deux stimuli dichotiques ; (ii) du type de patron accentuel lexical (par défaut vs. marqué) ; (iii) de la lexicalité de la cible et des stimuli (mots vs. pseudomots).

Nous avons fait varier le nombre de phonèmes partagés par des stimuli dichotiques CVCV selon trois degrés d'identité segmentale : les stimuli partagent soit seulement un phonème (la 2ème voyelle, V2) ; soit la 2ème syllabe ; soit la 1ère syllabe et V2, auquel cas ils ne diffèrent que par la 2ème consonne. De plus, nous avons construit des stimuli dichotiques dans lesquels le patron accentuel qui est susceptible de se combiner avec la chaîne phonémique correspond soit au patron par défaut, soit au patron marqué.

En portugais, le patron accentuel non marqué (ou par défaut) d'un mot bisyllabique se manifeste par la réalisation d'un accent sur la 1ère syllabe, alors que le patron accentuel marqué se manifeste par la réalisation d'un accent sur la 2ème syllabe ; cet accent est alors indiqué à l'écrit par un signe diacritique sur la 2ème syllabe : réalisation d'un accent aigu sur une voyelle et d'un tilde (~) sur la 1ère voyelle d'une diphtongue. Ainsi, à l'écrit, les graphies "am" et "ão" désignent la même diphtongue /ãw/, dans le premier cas inaccentuée et dans le second accentuée. Les syllabes accentuées sont réalisées avec une voyelle pleine, un allongement de la durée et une augmentation de l'intensité, alors que la plupart de syllabes inaccentuées sont réalisées avec une voyelle réduite /9/. Sur le plan de la structure métrique, les syllabes accentuées sont fortes et la plupart des syllabes inaccentuées sont faibles. Cependant, certaines syllabes sont intrinsèquement lourdes, ce qui signifie que la voyelle n'est jamais réduite en position inaccentuée ; de plus, dans un registre de langue soutenue, les voyelles peuvent ne pas être réduites quand elles sont inaccentuées. On observe par conséquent en portugais une distinction de poids syllabique qui est indépendante de la réalisation de l'accent /10/. L'objectif de cette étude étant de tester si le patron accentuel et la chaîne phonémique sont représentés séparément, le contraste accentuel ne doit être exprimé que dans la dimension prosodique. Par conséquent, nous avons choisi des groupes d'items dont les voyelles ne sont jamais réduites en position inaccentuée.

2. EXPÉRIENCE

Le but de l'expérimentation est de tester si, dans une situation d'écoute dichotique, les sujets peuvent avoir l'illusion d'entendre un mot ou un pseudomot-cible spécifié à l'avance, alors que l'information phonémique provient d'une oreille et l'information du patron accentuel de l'autre oreille. Prenons par exemple le mot portugais *CANsam* (les majuscules seront utilisées pour désigner une syllabe accentuée). Ce mot a les mêmes constituants phonémiques (mais pas le même patron accentuel) que le mot *canÇÃO* présenté à une oreille, et a le même patron accentuel (mais pas la même chaîne phonémique) que le mot *LImam* présenté à l'autre oreille (les graphies "am" et "ÃO" désignent la même diphtongue /ãw/, dans le premier cas inaccentuée et dans le second accentuée). La question est la suivante : le sujet perçoit-il la cible *CANsam* lorsqu'on lui présente auditivement la paire de stimuli dichotiques *canÇÃO-LImam*? Si dans cette situation, appelée *essai expérimental*, le sujet a parfois l'illusion d'entendre la cible, cela pourrait appuyer l'hypothèse que le patron accentuel et la chaîne phonémique sont représentés séparément et combinés de façon erronée.

La totalité des fausses détections de la cible *CANsam* observées à l'essai expérimental ne peut cependant pas être attribuée en toute certitude à une fausse combinaison du patron accentuel et de la chaîne phonémique. En raison de la perte de discriminabilité perceptive liée à la présentation dichotique, le sujet pourrait mal interpréter le patron accentuel du stimulus *canÇÃO*, qui a les mêmes constituants phonémiques que la cible *CANsam*. C'est pourquoi on construit un *essai contrôle*, où aucun des stimuli dichotiques n'a le patron accentuel de la cible : la cible *CANsam* a la même chaîne phonémique que le mot *canÇÃO* présenté à une oreille (comme dans l'essai expérimental), mais n'a pas le même patron accentuel que le mot *liMÃO* présenté à l'autre oreille. Par conséquent, afin d'établir l'occurrence de fausses combinaisons du patron accentuel et de la chaîne phonémique, le pourcentage de fausses détections de la cible *CANsam* doit être plus élevé aux essais expérimentaux qu'aux essais contrôles.

Le type d'essais décrits ci-dessus sont des *essais négatifs*, puisque la cible n'est pas présente dans la paire de stimuli dichotiques. Comme on peut le voir dans le tableau 1, qui présente l'ensemble complet des essais construits pour la cible *CANsam*, nous avons aussi construit deux *essais positifs* pour chaque cible, où la cible est présente dans la paire de stimuli. Ces essais positifs, expérimental et contrôle, sont construits de la même manière que les essais négatifs correspondants : par exemple pour la cible *CANsam*, le distracteur *LImam* apparaît dans les deux essais expérimentaux (négatif et positif) et le distracteur *liMÃO* dans les deux essais contrôles. Comme nous le verrons dans l'analyse

des résultats, l'utilisation d'essais négatifs et positifs permet de calculer un indice de discriminabilité (*d'*) qui est relativement indépendant des biais de réponse.

Tableau 1 : Essais pour le mot-cible "CANsam"

	Essai Expérimental	Essai Contrôle
Négatifs (cible absente)	canÇÃO - LImam	canÇÃO - liMÃO
Positifs (cible présente)	CANsam - LImam	CANsam - liMÃO

N. B. : "am" et "ÃO" désignent la même diphtongue /ãw/, inaccentuée et accentuée, respectivement

Méthode

- Matériels.

Le matériel a été construit à partir de doublets de paires accentuelles de cibles CVCV. Les membres d'une paire accentuelle se différencient seulement par la position de l'accent, qui porte sur la 1ère syllabe (Accent 1 : patron accentuel non marqué) ou sur la 2ème syllabe (Accent 2 : patron accentuel marqué). Ainsi, l'ensemble des essais présentés dans le tableau 1 a été construit pour le mot-cible *CANsam* à partir des deux paires accentuelles *CANsam/canÇÃO* et *LImam/liMÃO*. En suivant le même principe, nous avons construit trois autres ensembles d'essais, un pour chacune des trois cibles *canÇÃO*, *LImam* et *liMÃO*.

En plus de l'accent, nous avons manipulé deux autres facteurs susceptibles d'influencer les fausses combinaisons du patron accentuel et de la chaîne phonémique, à savoir la lexicalité et l'identité segmentale entre stimuli. Les cibles et les stimuli dichotiques sont soit des mots, soit des pseudomots (la lexicalité des cibles et des stimuli covarie). Les pseudomots-cibles ont été construites en changeant la 1ère consonne des mots-cibles (exemple : à la paire accentuelle de mots-cibles *CANsam/canÇÃO* correspond la paire de pseudomots-cibles *GANsam/ganÇÃO*). L'identité segmentale (IS) concerne le nombre de phonèmes que partagent les stimuli dichotiques ; ce nombre varie selon trois degrés : les stimuli partagent soit seulement V2 (IS1, comme dans l'exemple précédent *CANsam-LImam*) ; soit la 2ème syllabe (IS2, comme dans *VIram-FUram*) ; soit la 1ère syllabe et V2 (IS3, comme dans *PENdam-PENsam*).

Enfin, pour palier au petit nombre de paires accentuelles utilisables, nous avons construit trois groupes d'items, en essayant que les cibles soient associées aux trois conditions d'identité segmentale. Ainsi, par exemple pour l'essai expérimental négatif et pour la cible *CANsam*, nous avons comme stimuli dans les groupes d'items 1, 2 et 3, respectivement, les paires *canÇÃO-LImam* (IS1), *canÇÃO-PENsam* (IS2) et *canÇÃO-CANtam* (IS3). Au total, le test comporte dès lors 72 cibles, 36 mots et 36 pseudomots.

Les mots et pseudomots ont été lus par un homme de langue maternelle portugaise et enregistrés en chambre sourde. Nous avons utilisé l'algorithme SOLA de compression-expansion du signal afin d'égaliser la durée des items de chaque paire de stimuli dichotiques, sur base de leur durée moyenne. Le début et la fin de chaque paire a ensuite été synchronisé.

- Sujets et procédure

Soixante étudiants de langue maternelle portugaise de la Faculté de Psychologie de l'Université de Lisbonne ont été répartis en trois groupes de 20. Chaque groupe passait les essais correspondant à un des trois groupes d'items mots et aux pseudomots correspondant.

Les sujets ont passé le test deux-à-deux, en écoutant les stimuli avec un casque. Ils avaient pour tâche d'écouter chaque paire de stimuli et de répondre si oui (en inscrivant "+") ou non ("-") ils entendaient le mot- ou pseudomot-cible présenté sur un carnet. Ils entendaient une paire de stimuli toutes les 3 secondes. La moitié des sujets de chaque groupe passait les essais correspondant aux mots lors de la 1ère session, et ceux correspondant aux pseudomots lors de la seconde, l'ordre inverse étant utilisé pour l'autre moitié des sujets. Chaque session était précédée d'une phase d'entraînement. Lors de l'expérience, chaque essai négatif correspondant à une cible était présenté 6 fois et chaque essai positif 2 fois. La présentation à l'oreille gauche ou droite des stimuli de chaque paire dichotique était inversée lors des répétitions. Les essais étaient présentés dans un ordre pseudo-aléatoire.

3. RÉSULTATS

3.1. RÉSULTATS OBTENUS AUX ESSAIS NÉGATIFS (CIBLE ABSENTE)

Afin d'établir l'occurrence de fausses combinaisons du patron accentuel et de la chaîne phonémique, il faut vérifier que le nombre de fausses détections de la cible est plus élevé aux essais expérimentaux qu'aux essais contrôles. Afin d'évaluer et de comparer les fausses combinaisons dans les différentes conditions, nous avons calculé un taux de fausses détections en divisant le nombre de fausses détections obtenues aux essais expérimentaux par la somme des fausses détections obtenues aux essais expérimentaux et contrôles. Un taux supérieur à 0.5 signale ainsi la présence de fausses combinaisons.

Nous avons réalisé des analyses de variance par sujets (F1) et par items (c'est-à-dire par paires accentuelles, F2). Le plan expérimental comporte cinq facteurs : Essai (expérimental vs. contrôle), Accent (1 vs. 2), Lexicalité (cible et stimuli mots vs. pseudomots), Identité Segmentale (IS1, 2 ou 3), et Groupe d'items (3 groupes). Dans la description des résultats, nous ne retiendrons en règle générale que les effets qui sont significatifs ou marginalement significatifs à la fois dans l'analyse par sujets et dans l'analyse par items.

Le facteur Essai est significatif (F1 et F2 : $p < .0005$). L'interprétation de cet effet dépend cependant de la présence ou non d'interactions impliquant ce facteur.

L'influence du patron accentuel sur les fausses combinaisons est attestée par l'interaction significative entre le facteur Accent et le facteur Essai (F1 et F2 : $p < .0005$). Bien que le taux de fausses détections soit supérieur à 0.5 dans les deux conditions de patron accentuel (cf. tab. 2), ce taux est plus élevé lorsque la cible a un patron non marqué (Accent 1) que lorsque la cible a un patron marqué (Accent 2).

L'influence de l'identité segmentale sur les fausses combinaisons est attestée par l'interaction significative entre le facteur Identité Segmentale et le facteur Essai (F1 : $p < .0005$; F2 : $p < .025$). Bien que le taux de fausses détections soit supérieur à 0.5 dans les trois conditions d'IS (cf. tab. 3), les taux sont plus élevés lorsque les stimuli partagent au moins une syllabe (IS2 et IS3) que lorsqu'ils ne partagent qu'une voyelle (IS1). Par contre, le fait de partager une syllabe et une voyelle (IS3) plutôt que seulement une syllabe (IS2) ne semble pas pertinent.

La lexicalité de la cible et des stimuli ne semble pas avoir d'influence sur les fausses combinaisons, puisque les taux de fausses détections sont similaires dans les conditions mots et pseudomots (cf. tab. 4). Ainsi, bien que le facteur Lexicalité soit significatif (F1 : $p < .001$; F2 : $p < .01$), il n'interagit pas avec le facteur Essai.

Enfin, il faut signaler que nous observons un effet significatif du facteur Groupe d'items (F1 : $p < .05$; F2 : $p < .0005$), ainsi qu'une interaction significative entre ce facteur et le facteur Essai (F1 : $p < .0005$; F2 : $p < .01$). Aucune interaction double ni triple impliquant le facteur Essai n'est significative à la fois dans l'analyse par sujets et par items.

Tableau 2 : Pourcentages moyens et taux de fausses détections en fonction de l'accent (sur les 3 groupes d'items)

	Type d'essai		
	Expérimental	Contrôle	Taux
Accent 1	40.7%	21.9%	0.65
Accent 2	35.9%	29.4%	0.55

Tableau 3 : Pourcentages moyens et taux de fausses détections en fonction de l'identité segmentale (sur les 3 groupes d'items)

	Type d'essai		
	Expérimental	Contrôle	Taux
IS 1	36.0%	28.7%	0.55
IS 2	40.2%	23.8%	0.63
IS 3	38.6%	24.3%	0.62

Tableau 4 : Pourcentages moyens et taux de fausses détections en fonction de la lexicalité (sur les 3 groupes d'items)

	Type d'essai		
	Expérimental	Contrôle	Taux
Mots	35.0%	22.5%	0.61
Pseudomots	41.5%	28.8%	0.59

3.2. RÉSULTATS OBTENUS AUX ESSAIS POSITIFS (CIBLE PRÉSENTE)

En moyenne, les sujet détectent correctement la présence de la cible dans 79% des cas. Les analyses de variance réalisées sur les pourcentages moyens de détections correctes montrent que les facteurs significatifs sont l'Identité Segmentale (F1 : $p < .0005$; F2 : $p < .05$), le Groupe d'items (F1 : $p < .025$; F2 : $p = .05$), et le facteur Essai ((F1 et F2 : $p < .025$). Aucune interaction impliquant le facteur Essai n'est significative à la fois dans l'analyse par sujets et par items.

3.3. SCORES *d'* : RÉSULTATS

Nous avons calculé le score *d'* de la théorie de détection du signal /11/ parce qu'il est plus fiable qu'un score basé seulement sur les fausses détections. En effet, en prenant en compte à la fois les fausses détections (faites aux essais négatifs) et les détections correctes (faites aux essais positifs), le score *d'* fournit un indice de discriminabilité relativement indépendant des biais de réponse des sujets. Nous avons calculé un *d'* Expérimental (*d'E*) et un *d'* Contrôle (*d'C*). Sous l'hypothèse que les combinaisons du patron accentuel et de la chaîne phonémique reflètent la discriminabilité et pas seulement des biais de réponse, on s'attend à une situation de plus faible discriminabilité aux essais expérimentaux qu'aux essais contrôles (c'est-à-dire *d'E* plus petit que *d'C*). Afin de pouvoir facilement évaluer l'influence des différents facteurs manipulés, nous avons présenté dans les tableaux la valeur de *d'C-d'E*. Une différence *d'C-d'E* supérieure à 0 signale la présence de fausses combinaisons.

Les analyses de variance par sujets (F1) et par items (c'est-à-dire par paires accentuelles, F2) montrent que le facteur Essai est significatif (F1 : $p < .0005$; F2 : $p < .01$). Néanmoins, l'interprétation de cet effet dépendra de la présence ou non d'interactions impliquant ce facteur.

Seule l'interaction entre Essai et Accent est significative dans les deux analyses (F1 : $p < .0005$; F2: $p < .005$). Comme on peut l'observer dans le tableau 5, la différence *d'C-d'E* est plus élevée lorsque la cible a un patron non marqué (Accent 1) que lorsqu'elle a un patron marqué (Accent 2). Le facteur Essai n'est en fait significatif dans les deux analyses que pour l'Accent 1

(F1 et F2 : $p < .0005$). Pour l'Accent 2, le facteur Essai est marginalement significatif dans l'analyse par sujets (F1 : $p = .051$) et n'est pas significatif dans l'analyse par items.

En ce qui concerne l'influence de l'identité segmentale, bien qu'on observe que ce facteur soit significatif (F1 : $p < .0005$; F2 : $p < .05$) et qu'il y ait la même tendance que dans l'analyse des fausses détections (cf. tab. 6), l'interaction entre les facteurs Essai et Identité Segmentale n'est significative que dans l'analyse par sujets (F1 : $p < .05$).

Par ailleurs, comme on peut l'observer dans le tableau 7, bien que le facteur Lexicalité soit significatif (F1 : $p < .0005$; F2 : $p < .001$), il n'interagit pas avec le facteur Essai.

Enfin, relevons que le facteur Groupe d'items est significatif (F1 : $p = .01$; F2 : $p < .001$), mais n'interagit pas avec le facteur Essai.

En résumé, les fausses combinaisons du patron accentuel et de la chaîne phonémique apparaissent au niveau de la discriminabilité, mais ce phénomène n'a lieu que dans certains cas. En accord avec les résultats observés sur la base des fausses détections, on note l'influence significative du patron accentuel sur les erreurs de combinaison, ainsi que l'absence d'influence de la lexicalité. Par contre l'influence de l'identité segmentale et celle du groupe d'items ne semblent pas apparaître au niveau de la discriminabilité telle que celle-ci peut être mesurée par l'indice *d'*.

Tableau 5 : Moyennes des scores *d'* en fonction de l'accent (sur les 3 groupes d'items)

	d' E	d' C	d'C - d'E
Accent 1	1.60	2.20	0.58
Accent 2	1.87	2.03	0.16

Tableau 6 : Moyennes des scores *d'* en fonction de l'identité segmentale (sur les 3 groupes d'items)

	d' E	d' C	d' C - d'E
IS 1	2.10	2.29	0.19
IS 2	1.56	2.01	0.45
IS 3	1.54	2.01	0.47

Tableau 7 : Moyennes des scores *d'* en fonction de la lexicalité (sur les 3 groupes d'items)

	d' E	d' C	d' C - d'E
Mots	1.98	2.37	0.38
Pseudomots	1.49	1.84	0.35

4. DISCUSSION

Les résultats suggèrent que le sujet peut entendre illusoirement un mot- ou pseudomot-cible en combinant l'information phonémique présentée à une oreille avec l'information du patron accentuel présentée à l'autre oreille. Il semble donc qu'avant la reconnaissance de la cible, le patron accentuel et la chaîne phonémique sont représentés séparément, sinon ils ne pourraient se combiner de manière erronée. Ce résultat semble valider l'hypothèse, postulée par les théories phonologiques non-linéaires, que le patron accentuel et la chaîne phonémique à laquelle il est associé constituent des représentations séparées.

L'objectif de l'étude présentée ici était de vérifier si l'apparition de l'illusion est liée à des contraintes spécifiques d'association des constituants phonologiques. En effet, lors d'une expérience précédente /ʃ, s/, nous avions observé que le phénomène illusoire n'avait lieu que dans le cas où les stimuli dichotiques étaient des mots partageant la 2ème syllabe. Nous ne retrouvons pas dans la présente expérience cette influence de l'identité segmentale. Bien que les taux de fausses détections soient plus élevés lorsque les stimuli partagent au moins une syllabe que lorsqu'ils ne partagent qu'une voyelle, cet effet d'identité segmentale n'est pas confirmé par l'analyse des scores *d'*. On ne peut dès lors pas affirmer avec certitude que l'effet d'identité segmentale est dû à la discriminabilité plutôt qu'à des biais de réponse des sujets.

Les résultats des deux études ne sont cependant pas nécessairement contradictoires. En effet, dans l'expérience précédente nous comparions une situation dans laquelle des stimuli mots bisyllabiques partageaient la 2ème syllabe avec une situation dans laquelle les stimuli étaient des mots trisyllabiques qui ne partageaient aucun phonème. Par contre, dans la présente expérience, les stimuli partageaient toujours au moins un phonème, et nous avons comparé le degré d'identité segmentale. Il est dès lors probable que l'effet qui avait été observé auparavant était du type "tout ou rien" : l'occurrence du phénomène illusoire ne dépendrait pas de la nature ou du nombre de phonèmes partagés, mais bien du fait qu'il y ait ou non un phonème en commun. Cette hypothèse devrait être vérifiée par une comparaison systématique de la présence vs. absence totale d'identité segmentale.

Par ailleurs, deux observations suggèrent que le phénomène illusoire pourrait avoir lieu à un niveau relativement précoce de traitement. D'une part, les analyses menées sur la base du score *d'* montrent que l'illusion ne résulte pas seulement d'un biais de réponse des sujets. D'autre part, le fait que la lexicalité de la cible et des stimuli ne semble pas affecter le phénomène illusoire suggère que celui-ci prend place à un niveau de traitement prélexical.

Cependant, le type de patron accentuel qui est combiné avec une chaîne phonémique paraît jouer un rôle déterminant dans l'apparition des erreurs de combinaison. En effet, le patron accentuel et la chaîne phonémique se combinent beaucoup plus fréquemment

de manière illusoire lorsque le patron accentuel est non marqué, c'est-à-dire lorsque ce patron est celui qui apparaît par défaut en portugais. De plus, le phénomène illusoire n'apparaît pas de manière significative lorsque le patron accentuel susceptible de se combiner avec l'information phonémique est marqué. Ce dernier résultat suggère que les deux suites phonémiques ne peuvent pas "échanger" leurs patrons accentuels. Au contraire, il semblerait que ce soit systématiquement le patron accentuel non marqué qui "détruise" le patron marqué.

La prévalence du patron non marqué conforte l'hypothèse d'une représentation phonologique des suprasegments par sous-spécification des éléments qui sont non marqués dans la langue et par spécification des éléments marqués. Une telle hypothèse devrait cependant être vérifiée dans une langue présentant la distribution inverse des patrons accentuels marqué vs. non-marqué. En effet, en nous limitant à l'étude d'une seule langue, la caractéristique marqué vs. non-marqué se confond inévitablement avec la position syllabique (en l'occurrence, en portugais, la réalisation de l'accent porte sur la 1ère syllabe d'un mot bisyllabique pour le patron accentuel non marqué et sur la 2ème syllabe pour le patron marqué).

En conclusion, si nous tenons compte des deux expériences réalisées jusqu'ici en portugais, il semble que le phénomène de combinaison illusoire du patron accentuel et de la chaîne phonémique subisse au moins deux types de contraintes. D'une part, une contrainte d'association pourrait être que les stimuli doivent partager (au moins) un constituant phonémique pour que le patron accentuel de l'un puisse se combiner avec la chaîne phonémique de l'autre. D'autre part, une autre contrainte serait que le système perceptif tendrait à favoriser la perception des éléments non marqués dans la langue. Ces deux types de contraintes semblent limiter fortement la notion de "séparabilité" qu'on peut appliquer au cas du patron accentuel et de la chaîne phonémique, et méritent dès lors de plus amples investigations.

Remerciements : Cette étude a été financée par le *Programme Cognisciences* (CNRS) grâce à une bourse post-doctorale accordée au premier auteur. Le second auteur est Chercheur Qualifié du Fonds National de la Recherche Scientifique (*FNRS*) de Belgique. Des financements ont été aussi obtenus par le *Human Frontier Science Program* (projet : Processing consequences of contrasting language phonologies), par le *FNRS* (Loterie Nationale 93 ; convention n° 8.4505.93), et par une *Action de Recherche Concertée* (projet : Le traitement du langage dans les différentes modalités : Approches comparatives). Nous tenons aussi à remercier le laboratoire de Phonétique d'Aix-en-Provence, et tout particulièrement R. Espesser, pour nous avoir aidé dans la préparation des stimuli sur leur système d'analyse-synthèse avec l'algorithme SOLA, ainsi que l'Institut de Phonétique de l'Université Libre de Bruxelles pour les facilités d'enregistrement et la Faculté de Psychologie de l'Université de Lisbonne pour les possibilités de test, et en particulier A. Verhaeghe pour l'organisation des séances de test.

RÉFÉRENCES

/1/ Halle, M. & Vergnaud, J.-R. (1987). An Essay on Stress. Cambridge MA: MIT Press.

/2/ Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Oxford: Basil Blackwell.

/3/ Hombert J.-M. (1986). Word games: some implications for analysis of tone and other phonological constructs. In Ohala & Jaeger (eds.), *Experimental Phonology*. pp. 175-186.

/4/ Kolinsky, R. & Morais J. (1992). Représentations intermédiaires dans la reconnaissance de la parole : apports de la technique de création de mots illusoires. *Proc. of 19th Journées d'Etude sur la Parole*, Bruxelles, 19-22 Mai, 129-133.

/5/ Kolinsky, R., Morais J., & Cluytens, M. (sous presse) Intermediate representations in spoken word recognition: evidence from word illusions, *Journal of Memory & Language*.

/6/ Cutting, J. E. (1976). Auditory and linguistic processes in spech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**, 114-140.

/7/ Pasdeloup, V., Morais, J. & Kolinsky, R. (1993). Are stress and phonemic string processed separately? Evidence from speech illusions. *Proc. of 3rd Eurospeech*, Berlin, 21-23 Septembre.

/8/ Pasdeloup, V., Kolinsky, R. & Morais, J. Are lexical stress and phonemic string represented separately? Evidence from speech illusions (soumis).

/9/ Martins, M. R. (1982). *Aspects de l'accent en portugais*. Hambourg: Buske Verlag.

/10/ Carvalho, J. M. (1989) Phonological conditions on Portuguese clitic placement: on syntactic evidence for stress and rhythmical patterns, *Linguistics*, **27**, 405-436.

/11/ Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

L'EFFET DU DÉBIT DE PAROLE SUR LE RÔLE DU POINT D'UNICITÉ

MONIQUE RADEAU, PHILIPPE MOUSTY ET VALÉRIE PASDELOUP

UNIVERSITÉ LIBRE DE BRUXELLES

Résumé

Le modèle de la cohorte attribue un rôle particulier au début des mots dans la reconnaissance de la parole. Il prédit notamment qu'un mot serait reconnu à son point d'unicité (PU, le phonème à partir du début de la séquence où il ne reste qu'un seul mot dans la langue). Dans l'expérimentation perceptive présentée ici, nous avons étudié l'effet du débit sur le rôle du PU en français. Nous avons utilisé une tâche de classification de genre et avons comparé les temps de réponse à des mots français trisyllabiques à PU précoce et tardif lus à débit rapide, moyen et lent. Un effet du PU a été obtenu dans la condition de débit moyen. Cet effet augmentait de façon très significative dans la condition lente et disparaissait complètement dans la condition rapide. Des résultats comparables ont été obtenus à partir d'un corpus de parole de synthèse où seul le débit a été modifié au moyen d'un algorithme de compression-expansion du signal.

Dans notre corpus, le débit des mots lus en condition rapide correspondait en fait à celui observé dans des corpus de parole spontanée. Les résultats obtenus pour des mots isolés suggèrent donc que l'effet du PU est stratégique.

Mots-clés: *Point d'unicité, Débit.*

INTRODUCTION

Certains modèles de reconnaissance des mots parlés ont attribué un rôle spécial au début des mots dans le traitement de la parole (Marslen-Wilson & Welch, 1978; Cole & Jakimik, 1980). L'un des plus connus est celui de la cohorte de Marslen-Wilson (Marslen-Wilson & Welsh, 1978). Ce modèle, dans sa version initiale, considère la reconnaissance des mots parlés comme un processus parfaitement adapté à la nature séquentielle de l'information qu'il reçoit. Ce processus est supposé opérer non seulement en temps réel mais encore de façon optimalement efficace, n'extrayant que la quantité d'information juste nécessaire pour spécifier le mot de façon unique. La première phase du processus serait une phase d'activation: l'information sensorielle des quelques 150 à 200 premières ms du mot (ce qui correspond aux deux ou trois premiers phonèmes) activerait en parallèle la cohorte de tous les candidats lexicaux commençant par cette séquence particulière. La phase suivante consisterait en une réduction progressive de la cohorte initiale. Cette réduction obéirait à un principe catégoriel, "tout ou rien", un candidat étant désactivé et sortant de la cohorte à la première incompatibilité avec l'information sensorielle qui s'accumule. Un mot serait reconnu sans traitement ultérieur dès qu'il resterait le seul candidat possible de la cohorte initiale, en d'autres termes dès son point d'unicité (PU). Ainsi par exemple, la reconnaissance du mot CAMEMBERT, dont le PU est le /ɑ̃/, ne requèrerait que le traitement des quatre premiers phonèmes.

L'hypothèse selon laquelle le point de reconnaissance correspondrait au PU prédit que le temps qu'il faut pour identifier un mot ou pour détecter un non-mot sera constant à partir du PU pour les mots ou du point de déviation (PD) pour les non-mots (le point critique dans un non-mot étant le moment où les données ne sont plus compatibles avec aucun candidat lexical) et indépendant de la longueur de l'item. La pente de la fonction qui relie le temps d'identification d'un mot ou le temps de détection d'un non-mot au temps qui s'écoule du début de l'item jusqu'au PU ou au PD devrait donc être proche de 1.

Une conséquence d'un traitement optimal de ce type est qu'il rend compte de la segmentation du discours continu dans lequel il n'y a pas de séparation physique

entre mots successifs. Si les mots sont reconnus l'un à la suite de l'autre, la reconnaissance d'un mot devrait permettre en général la localisation du début du mot suivant.

Les données empiriques cependant, ne fournissent que des arguments mitigés en faveur du modèle de la cohorte. Ainsi, il ne semble pas qu'on reconnaisse les mots un à un. Une démonstration classique est celle de Pickett & Pollack (1963) et Pollack & Pickett (1964) qui ont utilisé des mots extraits du discours continu et ont présenté un, deux ou trois extraits successifs. Les auteurs ont observé une augmentation du pourcentage d'identification du mot initial au fur et à mesure de l'ajout de mots successifs, un mot isolé n'étant indentifié correctement que dans 50 % des cas environ. Plus récemment, Grosjean (1985) et Bard, Shillcock & Altmann (1988) ont obtenu des résultats analogues, en particulier pour les mots courts.

Bien que ces données soient incompatibles avec une procédure strictement gauche-droite, démarrant à chaque nouveau mot, elles n'invalident pas de façon définitive le statut spécial de l'information initiale des mots. Les effets du contexte subséquent pourraient se limiter aux mots qui ne sont pas définis de façon unitaire à leur terminaison mais constituent d'autres mots possibles avec une partie du contexte subséquent (ex.: la mousse tâche VS la moustache). Le problème ne se pose évidemment pas dans les études qui ont utilisé des mots isolés, c'est-à-dire, pour reprendre l'expression de Grosjean & Gee (1987), l'analogue parlé du mot écrit. Comme ce type de matériel est celui qui a été utilisé dans la plupart des travaux relatifs au rôle du PU dans les mots ou du PD dans les non-mots, les travaux en question réunissaient des conditions optimales à la mise en évidence du statut spécial des débuts de mots.

L'argument le plus fort à l'appui du modèle de la cohorte est l'obtention par Marslen-Wilson (1984) d'une pente proche de 1 (.90) entre les temps de décision à des non-mots et la position du PD mesurés à partir du début de la séquence. Toutefois, Taft & Hambly (1986) et Goodman & Huttenlocher (1988) testant eux aussi l'effet du PD dans les non-mots ont obtenu des résultats suggérant que les sujets ne donnaient pas leur réponse au PD mais beaucoup plus tard. Selon Goodman & Huttenlocher (1988), l'effet proche de 100 % obtenu par Marslen-Wilson (1984) serait dû à l'utilisation, comme items à PD précoce, de non-mots présentant en leur début une séquence de phonèmes illégale. Ce matériel permet en effet au sujet de détecter un non-mot sans consulter son lexique mental.

Les études relatives à l'effet de la position du PU dans les mots ne posent évidemment pas le même problème. L'obtention de réponses plus rapides pour les mots à PU précoce que tardif ne peut s'expliquer en termes de la présence dans le matériel de séquences phonotactiques impossibles dans la langue. Parmi les travaux menés, peu ont tenté d'estimer la contribution quantitative de la position du PU à la reconnaissance des mots. L'un d'eux, cependant, a été réalisé par Radeau, Mousty & Bertelson (1989) qui ont comparés des noms de la langue française à PU précoce (CAMEMBERT) ou tardif (CONFESSION) dans une tâche de classification en genre. Les noms de chaque série étaient moitié féminins et moitié masculins. Le sujet devait décider du genre du mot entendu le plus vite possible en poussant une clé réponse à deux choix. Selon les conditions expérimentales, les effets du PU représentaient 30 % à 53 % de l'intervalle de temps entre les positions moyennes du PU dans chaque série de mots. Utilisant le même matériel mais la tâche de répétition immédiate, Radeau & Morais (1990) ont obtenu un effet du PU beaucoup plus faible (8 %).

Les résultats de Radeau et al. (1989), bien qu'ils ne soient pas compatibles avec la notion de procédure optimale au coeur du modèle de la cohorte, constituent un argument important pour le rôle du PU dans la reconnaissance des mots parlés. La diminution de l'effet dans la tâche de répétition n'invalide pas réellement le modèle, si l'on suppose que répéter un mot entendu n'implique pas nécessairement l'accès au lexique mental.

Un aspect particulier des résultats de Radeau et al. (1989) est que l'effet du PU était plus grand lorsque les mots de chaque série étaient présentés en deux blocs homogènes séparés que lorsqu'ils étaient mélangés. L'effet de 53 % rapporté ci-dessus correspondait à une telle présentation. En présentation mixte, les effets variaient de 30 % à 37 %. La différence entre les deux types de présentation était entièrement due à une diminution des temps de réponse (TRs) pour les mots de la série précoce dans la condition homogène. Ce résultat suggérait que l'effet du PU est influencé en tout cas partiellement par des facteurs stratégiques. Par ailleurs, les stratégies ont plus de chance d'être efficaces rapidement pour les mots de la série précoce. Pour ceux à PU tardif, elles devraient donner lieu à un plus grand

nombre d'erreurs. C'est en fait la tendance qui a été observée.

Cette suggestion d'implication de stratégies dans nos données nous a donc amenés à reconsidérer notre matériel du point de vue de la durée des mots. Puisque le développement de stratégies prend du temps, la durée devenait une variable critique. Notre matériel consistait en mots trisyllabiques de 948 ms de durée moyenne. Cette durée correspondait donc à un débit de 3.2 syll/sec, ce qui est extrêmement lent comparé aux débits généralement rapportés dans la littérature.

Pour le français, une partie des données provient de travaux portant sur des corpus de texte lu. Ainsi, Bartkova & Sorin (1987) ont rapporté des débits de 4.2 syll/sec à 6.1 syll/sec (moyenne: 5.2 syll/sec) pour quatre des locuteurs les plus représentatifs. Pasdeloup (1990) a obtenu dans un corpus lu 254 ms et 135 ms en moyenne pour les syllabes portant un accent final et les syllabes inaccentuées, respectivement. En extrapolant aux mots de trois syllabes portant un accent final, la durée totale serait de 524 ms (5.7 syll/sec). Des mesures réalisées sur des corpus de conversation enregistrée à l'insu des locuteurs donnent des débits de 5.5 syll/sec à 6.0 syll/sec (Malécot, Johnston & Kizziar, 1972). Grosjean & Deschamps (1972, 1973) rapportent un débit de 5.3 syll/sec pour des interviews radiophoniques.

Le débit du locuteur des expériences de Radeau et al. (1989) était donc extrêmement lent. Ceci peut provenir de la situation artificielle dans laquelle on demande à quelqu'un de lire des mots isolés et d'articuler clairement. En général, dans les études relatives à l'effet du PU ou du PD dans des items isolés, la durée de ceux-ci n'est pas précisée. Dans un cas cependant, une estimation a été possible. Il s'agit de l'expérience de Goodman & Huttenlocher (1988) où le débit était de 3.6 syll/sec, donc ici encore beaucoup plus lent que ceux obtenus lors de la lecture de texte ou la conversation normale.

La possibilité que la durée anormalement longue de nos items ait favorisé le développement de stratégies nous a amenés à examiner l'influence du débit sur l'effet du PU. Si des stratégies sont impliquées dans l'effet du PU, la taille de l'effet devrait augmenter de la condition rapide à la condition lente. Dans l'expérience 1, nous avons comparé trois versions de notre matériel précédent lues à débit moyen, lent ou rapide.

EXPÉRIENCE 1

Méthode

Le matériel était celui utilisé par Radeau et al. (1989). Il consistait en deux séries de 34 noms trisyllabiques, moitié féminin et moitié masculin, à PU précoce (tombant dans la première ou la deuxième syllabe) et à PU tardif (tombant dans la troisième syllabe ou après la fin du mot). Les mots des deux séries étaient appariés en nombre de phonèmes et en fréquence d'usage.

L'enregistrement a été effectué en chambre sourde. Le locuteur avait pour tâche de parler le plus vite possible (débit rapide), normalement (débit moyen) et le plus lentement possible (débit lent). La durée des mots a été mesurée après numérisation. Les durées moyennes dans les trois conditions étaient respectivement de 535 ms, 833 ms et 1376 ms. Elles correspondent à des débits de 5.6 syll/sec, 3.6 syll/sec et 2.2 syll/sec. Les positions du PU à partir du début du mot ont été mesurées pour chaque item dans chaque condition de débit. Les différences moyennes entre les positions du PU dans les deux séries de mots étaient de 154 ms, 295 ms et 506 ms dans les conditions de débit rapide, moyen et lent respectivement.

Les items étaient présentés en listes mixtes. La tâche était celle de classification de genre. Quarante-huit sujets de langue maternelle française ont participé à l'expérience perceptive. Ils ont été répartis en trois groupes, chaque groupe étant affecté à l'une des trois conditions de débit.

Résultats

Comme le montre le Tableau 1, l'effet du PU est fortement affecté par le débit. L'avantage pour les mots à PU précoce se manifeste à nouveau dans la condition de débit moyen ($F1(1,15)=33.69$, $p<.001$; $F2(1,66)=5.59$, $p<.001$), il augmente dans la condition lente ($F1(1,15)=235.9$, $p<.001$; $F2(1,66)=22.19$, $p<.001$) et disparaît complètement dans la condition rapide (F1 et F2 proches de zéro). Comparés à la différence entre les positions du PU dans les deux séries de mots, les effets représentent 0 %, 25 % et 46% pour les débits rapide, moyen et lent respectivement. Les différences entre conditions sont toutes hautement significatives (lent-moyen: $F1(1,30)= 15.34$, $p<.001$; moyen-rapide: $F1(1,30)= 16.78$, $p<.001$; lent-rapide: $F1(1,30)= 70.95$, $p<.001$).

Tableau 1: Temps de réponse (TR; ms) et pourcentage d'erreurs moyens dans les trois conditions de débit.

Débit	PU précoce	PU tardif	Différence
Rapide			
TR:	1019	1021	-2
Erreurs:	3,5	5,3	1,8
Moyen			
TR:	1143	1217	74
Erreurs:	2,4	4,7	2,3
Lent			
TR:	1325	1557	232
Erreurs:	3,5	11,5	8,0

Les mots à PU tardif donnent lieu à plus d'erreurs que ceux à PU précoce, en particulier dans la condition lente où la différence est significative ($F1(1,15)=20.29$, $p<.001$; $F2(1,66)=8.94$, $p<.001$).

Discussion

Les variations de débit obtenues dans un corpus de parole naturelle entraînent des modifications sur le plan segmental et suprasegmental (entre autres, qualité vocalique, fréquence fondamentale et chronométrage syllabique). Afin de s'assurer que les résultats de l'expérimentation précédente n'étaient pas liés à des variations autres que celle du débit, nous avons réexaminé l'effet du PU en utilisant un corpus de parole de synthèse où seul le paramètre de débit a été modifié.

EXPÉRIENCE 2

Méthode

La durée des mots a été modifiée au moyen de l'algorithme de compression-expansion SOLA (Roucus & Wilgus, 1985) sur un système d'analyse-synthèse. Nous sommes partis des mots de parole naturelle de la condition de débit moyen de l'expérience précédente et avons modifié la durée de façon linéaire afin qu'elle corresponde à la durée des mots des conditions de débit rapide et lent de l'expérience précédente.

L'expérience perceptive a été menée sur deux groupes de 18 sujets qui ont participé à l'une des deux conditions de débit.

Résultats

Les résultats sont présentés dans le Tableau 2. Ils vont dans le même sens qu'en parole naturelle. A nouveau la condition lente donne lieu à un effet hautement significatif du PU sur les TRs ($F1(1,15)=125.58$, $p<.001$; $F2(1,66)=10.01$, $p<.001$). Dans la condition rapide, l'effet est significatif à l'analyse par sujets ($F1(1,15)=9.16$, $p<.001$) mais pas à l'analyse par items ($F2=1.08$). Les effets, comparés à la différence entre les positions du PU dans les deux séries de mots, représentent 13 % dans la conditions rapide et 26 % dans la condition lente. La différence entre les deux conditions est significative ($F1(1,30)=7.33$, $p<.025$).

Comme dans l'expérience 1, les erreurs sont plus nombreuses pour les mots à PU tardif que pour ceux à PU précoce. La différence est significative dans la condition lente ($F1(1,15)=8.56$, $p<.001$; $F2(1,66)=5.20$, $p<.03$).

Tableau 2: Temps de réponse (TR; ms) et pourcentage d'erreurs moyens dans les deux conditions de débit.

Débit	PU précoce	PU tardif	Différence
Rapide			
TR:	1007	1031	24
Erreurs:	4,0	7,2	3,2
Lent			
TR:	1461	1586	125
Erreurs:	5,3	9,6	4,3

CONCLUSIONS

Il apparaît que l'avantage associé à des mots à PU précoce s'accroît lorsque le débit ralentit. En fait, il n'y a pas d'effet du tout en condition de débit rapide. Cependant, le débit de la condition rapide (5.6 syll/sec) est celui qui se rapproche le plus de celui observé en lecture de texte ou en conversation spontanée (5.2 syll/sec à 6 syll/sec). Nos résultats montrent donc clairement que la position du PU n'influence pas la reconnaissance des mots parlés dans des conditions habituelles de conversation. Ceci est vrai même lorsqu'on utilise, comme c'était le cas dans nos expériences, des mots isolés qui constituent une

condition particulièrement favorable à l'utilisation du PU. L'intervention de composantes stratégiques est très probablement responsable de l'augmentation de l'effet à débit lent. Plus la durée des mots est longue, plus les stratégies ont le temps de se développer, d'où l'accroissement progressif de l'effet avec la durée lorsque l'on passe de l'une à l'autre des conditions de nos expériences, et l'augmentation des erreurs pour les mots à PU tardif.

On peut s'interroger sur la validité écologique d'une procédure de reconnaissance utilisant la position du point d'unicité. Des statistiques lexicales menées récemment en français (Mousty & Radeau, en préparation) sur les entrées de notre base de données BRULEX (Content, Mousty & Radeau, 1990) ont montré que les mots à PU précoce sont rares. La probabilité qu'un mot diverge des autres avant le dernier phonème n'est que de .42. Cette probabilité passe à .19, .10 et .03 pour des PU tombant deux, trois ou quatre phonèmes avant la fin du mot. Les mots qui ont leur PU avant la fin sont essentiellement des mots longs (de plus de six phonèmes) et ils sont peu fréquents. Ces tendances confirment celles observées par Luce (1986) pour l'anglais.

On pourrait se demander ce qu'il en est des enfants qui ont un lexique plus réduit que celui des adultes et donc des mots à PU plus précoce. En fait, il semble qu'ils n'utilisent pas non plus la position du PU (Charles-Luce & Luce, 1984). Si on accepte l'idée que le développement de stratégies s'accroît probablement avec l'âge, ce résultat est en accord avec une interprétation de l'effet du PU en termes de stratégies. Les conditions de laboratoire où des temps de réaction sont enregistrés constituent évidemment des cas particulièrement favorables à l'utilisation de telles stratégies.

RÉFÉRENCES

Bard, E.G., Shillcock, R.C., & Altmann, G.T.M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception and Psychophysics, 44*, 395-408.

Bartkova, K., & Sorin, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication, 6*, 245-260.

Cole, R.A., & Jakimik, J. (1980). A model of speech perception. In R.A. Cole (Ed.), *Perception and Production of Fluent Speech* (pp. 133-163). Hillsdale, NJ.: Lawrence Erlbaum.

Content, A., Mousty, P., & Radeau, M. (1990). BRULEX: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique, 90*, 551-566.

Goodman, J.C., & Huttenlocher, J. (1988). Do we know how people identify spoken words ? *Journal of Memory and Language, 27*, 684-698.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics, 38*, 299-310.

Grosjean, F., & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica, 26*, 129-156.

Grosjean, F., & Deschamps, A. (1973). Analyse des variables temporelles du français spontané. II. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique). *Phonetica, 28*, 191-226.

Grosjean, F., & Gee, J.P. (1987). Prosodic structure and spoken word recognition. *Cognition, 25*, 135-155.

Luce, P.A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics, 39*, 155-158.

Luce, J.A., & Luce, P.A. (1984). Some structural properties of words in young children's lexicons. *Research on Speech Perception*, Progress Report n° 10, 323-344.

Malécot, A., Johnston, R., & Kiziar, P.-A. (1972). Syllabic rate and utterance length in french. *Phonetica, 26*, 235-251.

Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. In : H.S. Bouma & D.G. Bouwhuis (Eds.), *Attention and Performance X: Control of language processes* (pp. 125-150). Hillsdale, NJ : Lawrence Erlbaum.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Mousty, P., & Radeau, M. (en préparation). Une analyse computationnelle des points d'unicité et de l'évolution temporelle des compétiteurs dans le traitement en ligne des mots parlés en français.

Pasdeloup, V. (1990). Modèle de règles rythmiques du français appliqué à la synthèse de la parole. Unpublished Doctoral thesis. Université de Provence.

Pickett, J.M., & I. Pollack. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech, 6*, 151-164.

Pollack, I., & Pickett, J.M. (1964). Intelligibility of excerpts from fluent speech: Auditory vs. structural

context. *Journal of Verbal Learning and Verbal Behavior, 3*, 79-84.

Radeau, M., & Morais, J. (1990). The uniquenes point effect in the shadowing of spoken words. *Speech Communication, 9*, 155-164.

Radeau, M., Mousty, P., & Bertelson, P. (1989). The effect of the uniqueness point in spoken-word recognition. *Psychological Research, 51*, 123-128.

Roucus, S., & Wilgus, A.M. (1985). High quality time-scale modification for speech. *ICASSP*, TAMPA, 493-496.

Taft, M., & Hambly, G. (1986). Exploring the cohort model of spoken word recognition. *Cognition, 22*, 259-282.

Remerciements

Le premier auteur est Maître de Recherches au FNRS. Le travail a bénéficié de subventions du Ministère de l'Éducation de la Communauté Française de Belgique (Action de Recherches concertées "Traitement du Langage dans les Différentes Modalités: Approche comparative"). Nous remercions le Laboratoire de Phonétique d'Aix-en-Provence, et en particulier Robert Espesser qui nous a aidé dans la préparation de nos stimuli sur le système d'analyse-synthèse avec l'algorithme SOLA.

SEGMENTATION ET RECONNAISSANCE DES MOTS PARLÉS EN CONTEXTE : UNE APPROCHE COMPUTATIONNELLE *

ALAIN CONTENT ET PASCAL STERNON

LABORATOIRE DE PSYCHOLOGIE EXPÉRIMENTALE
UNIVERSITÉ LIBRE DE BRUXELLES

Résumé

Nous présentons une approche de la reconnaissance des mots parlés en contexte basée sur les techniques d'apprentissage connexionnistes, qui utilise une variante nouvelle de réseau récurrent simple. Des simulations basées sur de petits lexiques montrent que le système fournit une solution optimale à la double contrainte de *rapidité*, qui exige que le traitement lexical soit immédiat, et de *fiabilité*, qui impose de différer les décisions d'identification. Contrairement à un argument souvent avancé, nos simulations indiquent que l'existence de mots enchâssés dans d'autres mots n'est pas incompatible avec la notion de traitement immédiat et continu.

Mots-clés: *lexique, parole continue, segmentation , connexionnisme, réseaux récurrents.*

INTRODUCTION

Jusqu'aux années récentes, la plupart des travaux expérimentaux visant à explorer la nature des mécanismes de reconnaissance des mots parlés chez le sujet humain se sont limités à considérer le traitement de mots isolés. Cette tendance s'explique probablement en partie par l'influence sur l'évolution de la psycholinguistique des recherches utilisant le support écrit, plus nombreuses et techniquement plus faciles à mener (Frauenfelder et Tyler, 1987).

Par rapport à la modalité visuelle, la parole présente deux caractéristiques cruciales, qui ont souvent été notées, et qui justifient la mise en oeuvre de processus distincts : (a) le signal de parole se déroule dans le temps; et (b) les frontières entre unités linguistiques ne sont pas explicitement marquées dans le signal. Nous examinerons ici les conséquences de ces deux caractéristiques pour la compréhension des mécanismes d'appariement entre le signal de parole et les représentations lexicales.

La première caractéristique soulève le problème de la nature immédiate ou différée des opérations de comparaison entre l'information sensorielle et les représentations lexicales. Dans les théories qui supposent que l'appariement est *immédiat*, on admet que chaque événement acoustique-phonétique détermine une nouvelle comparaison. L'avantage computationnel présumé est la rapidité, puisque l'accès à l'information lexicale est possible dès que le donné sensoriel suffisant pour isoler un candidat est disponible (Marslen-Wilson et Welsh, 1978; Marslen-Wilson, 1987). La difficulté essentielle consiste à spécifier la nature des contraintes qui guident l'appariement, de manière à éviter d'entretenir toutes les hypothèses compatibles avec n'importe quel segment de signal (Shillcock, 1990).

Par exemple, dans le modèle bien connu de la Cohorte (Marslen-Wilson et Welsh, 1978), seuls les candidats lexicaux alignés à partir d'une frontière de mot sont pris en considération. Cette hypothèse présuppose que la fin du mot précédent ou le début d'un nouveau mot peut être identifié sans incertitude à partir du signal. Or, du fait de l'existence de mots *enchâssés* (ex. "cerf" et "serpent"), il n'est pas toujours possible de prédire la frontière lexicale avant le début du mot suivant. Comme le montre l'exemple ci-dessous, l'analyse appropriée peut dépendre d'informations beaucoup plus tardives dans la phrase.

(1) ʒevyœsɛʀpɑ̃dɑ̃sɑ̃tɑ̃to

(2) ʒevyœsɛʀpɑ̃dɑ̃sɑ̃tɛt

On sait qu'il existe des indices phonétiques, prosodiques, phonotactiques et phonologiques qui indiquent des frontières potentielles ou même probables (Church, 1987). Ainsi, la présence d'un accent principal sur la 5ème syllabe dans (1) et la 4ème dans (2) suffit probablement à lever l'ambiguïté. Il semble cependant généralement admis que ces indices ne sont pas suffisants pour établir une analyse lexicale correcte et non-ambiguë. Des résultats expérimentaux récents (Tabossi, 1993) suggèrent que les mots enchâssés peuvent être activés malgré la présence d'indices de segmentation clairs. Il est donc peu probable qu'une solution *complète* du problème de la segmentation lexicale résulte d'heuristiques liées à des caractéristiques du signal. L'autre

* *Les recherches décrites ont été menées dans le cadre d'une Action de Recherche Concertée soutenue par le Ministère de l'Education de la Communauté française de Belgique (projet intitulé "Language Processing in Different Modalities: Comparative Approaches"). Nous remercions Jeff Elman qui nous a fourni le simulateur TLearn, et Uli Frauenfelder pour ses commentaires stimulants.*

possibilité, garantissant la fiabilité suppose que les décisions sont retardées jusqu'à ce qu'une quantité d'information suffisante soit disponible pour éliminer toute ambiguïté (Cf. Grosjean, 1985).

Comme l'ont montré des analyses détaillées de son comportement (Frauenfelder et Peeters, 1990, 1993), le modèle TRACE, développé par McClelland et Elman (1986) constitue une illustration d'un mécanisme qui réalise un compromis optimal entre ces deux exigences opposées de rapidité et de fiabilité. Le problème de la segmentation lexicale est résolu dans TRACE par la combinaison de trois sources de contraintes: la réception séquentielle de l'information sensorielle, qui s'accumule progressivement au niveau de la "trace" constituée par l'ensemble des détecteurs de traits et de phonèmes dupliqués pour chaque instant de la plage temporelle considérée; l'appariement entre la trace et l'information lexicale, qui est basé sur un principe de comparaison exhaustive et graduelle; et la compétition directe entre hypothèses lexicales, en proportion de leur degré de superposition temporelle. La réception séquentielle de l'information assure que les hypothèses lexicales compatibles avec le signal seront activées dès que possible. De plus, l'existence d'une trace sensorielle couvrant une plage temporelle de signal permet en principe au système de modifier la force des hypothèses lexicales en prenant en considération l'ensemble de l'information disponible à chaque instant.

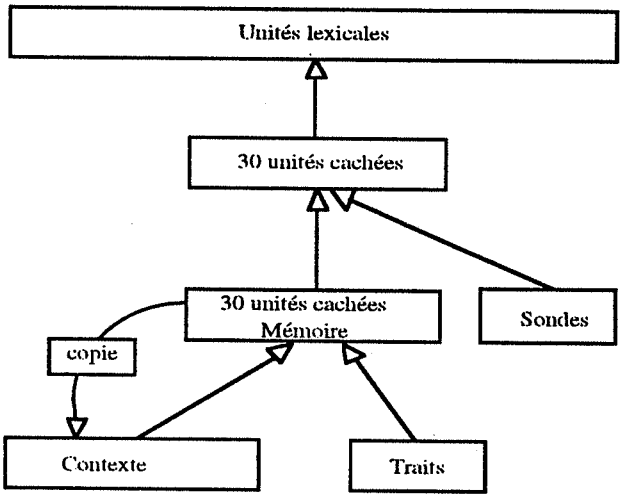
TRACE a cependant été critiqué à cause du mode de traitement et de représentation de l'information temporelle adopté. Le principe de réduplication spatiale de l'information est à la fois peu efficient et peu plausible, puisqu'il entraîne une multiplication des unités lexicales (Norris, 1990), et peu satisfaisant sur le plan computationnel, parce qu'il n'offre aucune flexibilité par rapport aux variations temporelles du signal.

Une solution potentielle consiste à utiliser des réseaux à topologie récurrente, auxquels peuvent être appliqués des algorithmes d'apprentissage automatique. Ces systèmes présentent deux avantages. D'une part, l'information temporelle n'est pas représentée explicitement, mais implicitement du fait de la topologie même des connexions dans le réseau. Par exemple, dans les *réseaux récurrents simples (SRN)* proposés par Elman (1990), la couche intermédiaire du réseau reçoit à chaque cycle non seulement un nouvel élément externe, mais également son propre état au cycle précédent, qui constitue donc une représentation compacte du contexte temporel. Cette *boucle* permet aux unités de la couche intermédiaire de fonctionner comme une mémoire, en combinant à chaque cycle leur état précédent avec l'information nouvelle. D'autre part, contrairement à des systèmes comme TRACE, dans lesquels les connexions doivent être fixées par le concepteur, ces réseaux peuvent être soumis à une procédure d'apprentissage (par exemple l'algorithme de rétro-propagation de l'erreur; Rumelhart, Hinton et Williams, 1986). La nature de la mémoire développée par le système sera contrainte par la tâche à laquelle le réseau est soumis, puisque c'est la réponse attendue qui détermine en dernier recours l'état des unités intermédiaires, et le mode de combinaison de l'état précédent avec le contexte.

Une tentative de simulation de la reconnaissance des mots parlés à l'aide d'un SRN a été réalisée par Norris (1990). Norris a utilisé un lexique de 50 formes et a entraîné un SRN sur une séquence continue de mots, présentés segment par segment. A chaque mot correspondait une unité de sortie particulière, et la réponse attendue consistait donc à activer l'unité lexicale appropriée. Malgré des résultats positifs et des similitudes intéressantes avec le modèle de la Cohorte, Norris conclut de manière négative, en observant que le SRN est incapable de réviser ses choix en fonction de l'information ultérieure, et qu'il ne peut donc résoudre le problème de la reconnaissance de mots enchâssés.

DESCRIPTION DU RÉSEAU

Le but de notre étude est de poursuivre et d'étendre la démarche entamée par Norris, en examinant une nouvelle variante de réseau récurrent qui est susceptible de résoudre le problème du contexte "à droite". L'idée essentielle consiste à modifier la tâche imposée au réseau de manière à ce que son état reflète non seulement les hypothèses sur le mot en cours, mais également sur le mot immédiatement précédent.



La figure 1 présente la topologie du réseau. Comme dans le cas du SRN, le signal d'entrée est constitué de la description simplifiée (sous forme de 7 traits repris de McClelland et Elman, 1986) d'un segment de parole, auquel vient s'adjoindre la copie de l'état d'activation des unités "Mémoire" au cycle immédiatement précédent. La nouveauté que nous avons introduite consiste en l'ajout d'un groupe d'unités "Sondes", qui constituent un signal externe supplémentaire, et détermine la nature de la réponse (Cf. St John et McClelland, 1990 pour une architecture similaire appliquée au traitement de phrases). La sonde et l'état des unités-mémoire sont combinées dans une seconde couche. Les unités de réponse sont des détecteurs de mots.

MODALITÉS D'ENTRAÎNEMENT

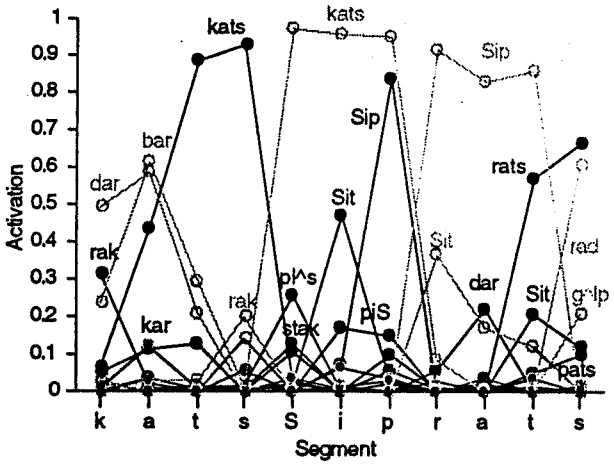
Quatre simulations ont été menées. Chacune était basée sur un lexique de 20 formes, et utilisait les mêmes paramètres (vitesse d'apprentissage=.05, momentum=0). Le premier lexique (Lex1) comprenait 10 formes de 3 ou 4 segments et leur inverse, suivant une procédure

adoptée précédemment par Norris (1990) pour vérifier l'encodage de l'ordre des événements. Lex2 utilisait le même ensemble, mais 4 mots ont été modifiés en supprimant le segment final pour créer des conditions d'enchâssement initial ("sta" vs "star"). Lex3 et Lex4 sont basés sur un nouvel échantillon de mots, dont 4 ont été modifiés de manière à créer respectivement une situation d'enchâssement central ("abi", "rabit") et final ("irg", "kirg")[1]. Chaque période d'entraînement consistait en une nouvelle séquence de 400 mots tirés aléatoirement, avec la restriction que 10% des paires possibles n'apparaissaient jamais. Ces combinaisons ont été réservées pour examiner les capacités de généralisation du réseau. Tous les mots avaient des probabilités d'occurrence égales.

Deux valeurs de sondes ont été utilisées. Pour la valeur *C*, la réponse attendue correspondait au mot en cours de présentation. Pour la valeur *P*, la réponse correspondait au mot immédiatement précédent. A chaque cycle, l'une des deux sondes était sélectionnée au hasard, de sorte que le réseau était confronté à un mélange aléatoire des deux tâches.

UN EXEMPLE

La figure 2 présente l'activation des 20 unités lexicales sur une séquence continue constituée par la suite des mots "kats", "ship" et "rats" (Lex1).

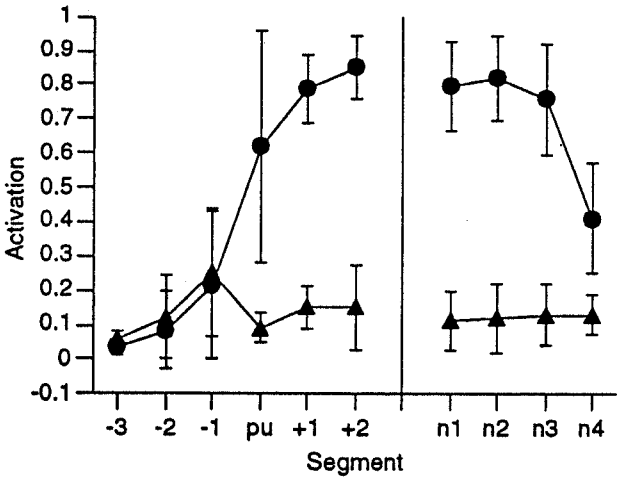


La séquence a été présentée deux fois, l'une avec la sonde C (lignes noires) et l'autre avec la sonde P (lignes grises). On peut constater que quelques candidats seulement sont actifs à chaque cycle. Comme l'avait montré Norris (1990), les membres des cohortes sont activés durant la présentation, mais d'autres candidats apparaissent également. Les deux tâches se complètent de manière précise. L'activation des unités en réponse à la sonde C —les hypothèses du réseau sur le mot en cours— se prolongent sans discontinuité durant la présentation du mot suivant du fait de la sonde P, de sorte que les mots identifiés conservent un haut niveau d'activation durant plusieurs cycles.

TRAITEMENT DU MOT EN COURS

Pour analyser la performance de manière plus systématique, une série de séquences comprenant toutes les combinaisons possibles de trois mots a été construite. Les analyses principales concernent la réponse à la sonde C durant la présentation du 2ème mot de chaque série, et les réponses à la sonde P durant la présentation du 3ème mot. Les réponses au premier mot n'ont pas été analysées en détail, parce qu'elles sont fortement influencées par l'état initial du réseau et ne reflètent donc pas correctement son régime normal de fonctionnement. Sauf mention, les résultats présentés concernent Lex1. Les résultats obtenus dans les autres cas sont généralement analogues.

La figure 3 présente l'état d'activation de l'unité-cible (correspondant au mot présenté) et du concurrent (le candidat incorrect le plus actif) à chaque cycle. La partie gauche correspond à la sonde C, et on a calculé l'activation moyenne en alignant sur le point d'unicité (PU) de la cible, c'est-à-dire le segment pour lequel la séquence présentée est compatible avec un seul candidat. Les barres verticales indiquent l'écart-type (par rapport à l'ensemble des 20 mots).



En moyenne, l'activation de la cible se différencie de celle des autres candidats à partir du PU du mot en cours. Le réseau se comporte donc de manière similaire au modèle de la Cohorte (Marslen-Wilson, 1990), malgré que la variabilité au cycle correspondant au PU soit importante, et que l'activation continue à croître durant les cycles suivants. Pour évaluer de manière plus fine la relation entre l'évolution de l'activation et le PU, nous avons calculé le *point d'isolation* de chaque mot, c'est-à-dire le premier cycle au cours duquel l'activation de l'unité cible dépasse celle de tous les autres candidats d'une valeur égale ou supérieure à 0.5 (Cf. Marslen-Wilson, 1990, et Bard, 1990, pour des suggestions analogues avec TRACE).

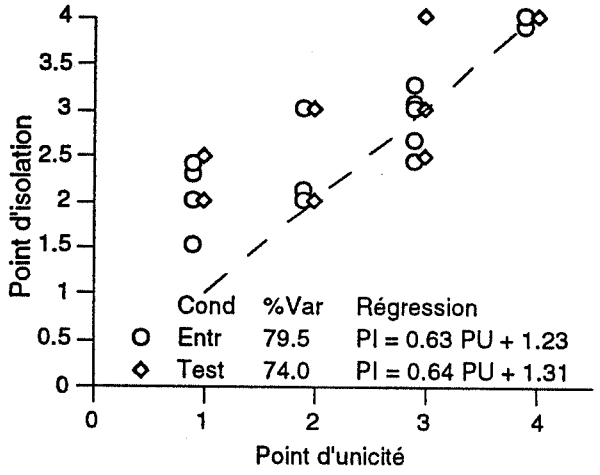
Le Tableau 1 présente les taux moyens d'isolation pour les 4 simulations. La performance plus faible pour Lex2 est due aux 4 mots enchâssés, pour lesquels le taux d'isolation en cours est nul. On constate également que les pertes de performance pour des séquences nouvelles sont faibles. Contrairement à ce que Norris

[1]Du fait de ces manipulations, les "lexiques" combinent des mots anglais, repris du corpus de TRACE, et des formes idiosyncratiques. Les formes lexicales sont notées entre guillemets, les séquences présentées entre barres. Nous utilisons les conventions de notation de TRACE (/S/ comme dans "ship", /^/ comme dans "plug").

(1990) concluait, il ne semble pas nécessaire d'entraîner le réseau sur toutes les séquences de mots possibles.

Corpus	Entraînement	Test
Lex1	87.2	75.0
Lex2	77.8	75.0
Lex3	95.3	97.5
Lex4	95.8	95.0

La figure 4 montre la relation entre le point d'unicité et le point d'isolation. La ligne hachurée indique la prédiction théorique du modèle de la cohorte (pente=1). Comme les analyses précédentes le laissaient entrevoir, il existe une dépendance étroite entre point d'isolation et point d'unicité. Il est néanmoins intéressant de constater que l'isolation est souvent plus tardive que la position du PU ne laisserait prédire, et ce décalage est particulièrement marqué pour les positions précoces. Il en résulte que si le comportement du réseau est analogue à celui du modèle de la Cohorte, il n'est cependant pas strictement équivalent.



Nous avons également examiné la nature des candidats lexicaux les plus activés. A chaque cycle, le concurrent a été identifié, pour autant que son niveau d'activation dépasse 0.25. Les concurrents ont ensuite été classés comme membre d'une cohorte, inversion, ou rime par rapport à la cible. Entre 50% et 90% des observations (selon les lexiques) peuvent être classées comme des membres des cohortes. Dans une faible proportion de cas apparaissent des candidats partageant la rime (entre 1 et 15%), ou correspondant à une inversion (2 à 9%). La faible proportion de concurrents correspondant à la séquence inversée montre que le réseau encode l'identité des segments et leur ordre d'apparition.

Globalement, tous les mots qui peuvent être identifiés avant leur fin sont effectivement isolés, et le point d'isolation correspond approximativement au point d'unicité lexical. Les hypothèses concurrentes correspondent le plus fréquemment à des membres de la cohorte. Ce comportement est analogue à celui qu'on observe dans la situation de dévoilement progressif (Grosjean, 1980).

TRAITEMENT DU MOT PRÉCÉDENT

L'analyse porte ici (pour chaque lexique) sur la réponse à la sonde P durant les cycles correspondant au 3ème mot, pour toutes les séquences de 3 mots possibles. Comme l'indique la figure 2 (partie de droite), l'activation de l'unité correspondant au mot précédent se maintient à un niveau élevé pendant toute la durée du mot en cours. Typiquement, l'activation tend à diminuer lors des derniers cycles. Ce fait s'explique probablement par les différences de longueur entre les mots. D'une manière générale, on le sait, la performance des réseaux est très sensible à la fréquence des essais d'apprentissage. Or tous les successeurs de longueur $\geq l$ contribuent à l'apprentissage de la tâche P durant les l premiers cycles, mais seuls ceux de longueur $> l$ fournissent des expériences d'apprentissage pour les cycles ultérieurs. L'activation correspondant au mot précédent aura donc naturellement tendance à s'éteindre graduellement.

Nous avons calculé la proportion de séquences pour lesquelles l'unité cible est effectivement plus activée (critère= 0.5) que toutes les autres. Comme pour la tâche C, le taux d'isolation voisine autour de 90%, qu'il s'agisse des séquences apprises ou des séquences de test. En particulier, les mots enchâssés au début (Lex2) qui ne pouvaient être isolés en cours de présentation donnent lieu à des taux d'isolation identiques aux autres mots.

Corpus	Séquences d'entraînement		Séquences de test	
	%Isol.	PI	%Isol.	PI
Lex1	87.2	1.26	80.1	1.11
Lex2	95.3	1.18	91.5	1.21
Lex3	95.6	1.18	91.4	1.21
Lex4	93.5	1.12	89.6	1.07

Sauf dans les cas de séquences ambiguës, qui sont examinées en détail ci-dessous, le point d'isolation devrait correspondre au premier segment lorsque les mots ont été isolés au cours de leur présentation. Les résultats montrent que la performance du réseau est effectivement très proche de cette valeur. Étant donne que la longueur des mots varie (de 2 à 5 segments dans le cas le plus extrême), cette analyse implique que le réseau prend effectivement en compte les frontières des mots pour déterminer la nature du mot précédent.

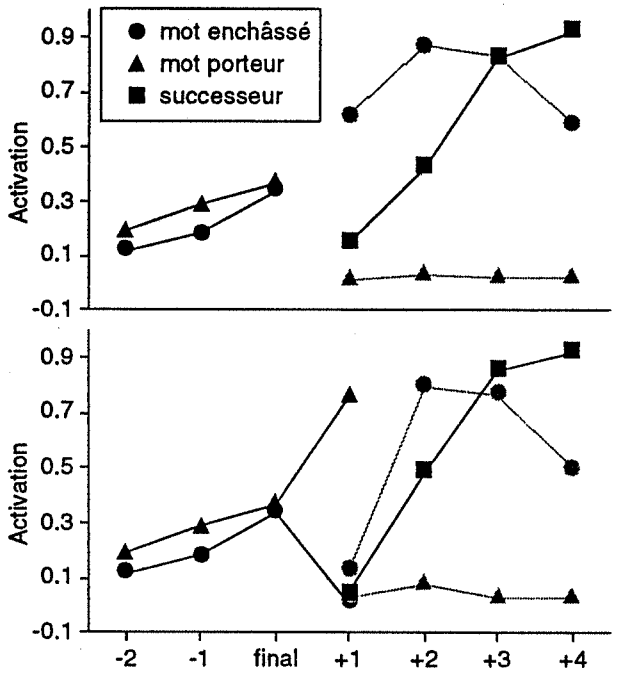
LES SÉQUENCES AMBIGUËS

Étant donné la nature des lexiques utilisés, les seules situations où une analyse immédiate ne permet pas de prendre une décision univoque sont celles où le mot présenté constitue le début d'un autre mot plus long (Lex2). On a vu que c'est effectivement dans ce cas que le réseau ne parvient pas à isoler le mot présenté avant le début du mot suivant.

Il reste que le moment où l'ambiguïté disparaît peut varier selon la nature du successeur. Par exemple, si le mot "sta" est suivi de "kats", l'ambiguïté est levée au premier segment (/k/) parce que le lexique n'inclut aucun mot compatible avec la séquence /stak/). Par contre, si le successeur est "pl^s" ou "rats", l'ambiguïté est pro-

longée dans la mesure où "stap" et "star" sont des mots du lexique utilisé. La comparaison de ces deux situations permet d'attester que la réponse du réseau est circonstancielle et que l'utilisation de la tâche P est de nature à résoudre les problèmes d'ambiguïté.

Le panneau supérieur de la figure 5 concerne l'ensemble des cas dans lesquels un mot enchâssé est suivi d'un autre mot qui élimine l'ambiguïté (Ex. /stakats/). Comme les mots enchâssés varient en longueur, ils ont été alignés sur leur segment final. Durant sa présentation, le mot inclus ("sta") ne se différencie pas des porteurs ("star", "stap"). Immédiatement après le dernier segment du mot enchâssé, la réponse à la tâche P fait apparaître que le réseau retient effectivement le mot enchâssé, et non le porteur, et cette réponse se maintient pendant toute la durée du mot suivant.



Le panneau inférieur reprend toutes les séquences dans lesquelles le premier segment du successeur est compatible avec un mot porteur (Ex. /starats/). On constate que sur le segment ambigu, le réseau isole (incorrectement) le mot porteur. Néanmoins, dès le cycle suivant, le mot enchâssé est clairement activé. Le résultat net est que l'ambiguïté prolongée de la séquence conduit à une activation transitoire du mot porteur, à un décalage d'un cycle de l'isolation du mot enchâssé, et à une différence légère dans l'activation initiale du successeur.

LIMITES

Une première limite du travail concerne la taille du corpus utilisé. Nous avons opté pour des simulations de taille minimale, qui présentaient l'avantage de permettre une analyse fouillée des performances du réseau. Des simulations sont en cours avec des lexiques plus grands, mais plusieurs difficultés, liées à la fois aux ressources computationnelles disponibles, à la nature de l'algorithme d'apprentissage itératif utilisé, et au nombre d'unités lexicales, empêchent de passer à un ordre de grandeur comparable à des lexiques naturels.

Une autre limitation consiste à utiliser une description du signal discrétisée et simplifiée à l'extrême. Ce choix reflète évidemment ici un principe méthodologique et non une option théorique. Au contraire, une piste intéressante pour la poursuite de cette recherche concerne l'utilisation de représentations d'entrée plus riches et d'un grain plus fin. Des travaux récents (Norris, 1990; Abu-Bakar, 1993) suggèrent que la performance du SRN n'est pas complètement perturbée par des variations de durée des séquences ou de certains éléments de celles-ci. Cette flexibilité temporelle est susceptible de constituer un avantage important par rapport à d'autres architectures, comme celle du modèle TRACE, notamment.

DISCUSSION

Malgré ces limitations, nous pensons que le modèle ouvre des perspectives intéressantes. Si on se réfère aux distinctions proposées par Frauenfelder (1991), on peut caractériser les mécanismes d'appariement mis en oeuvre par le réseau de la manière suivante: (1) il s'agit d'un processus d'activation directe; l'entrée est mise en relation simultanément avec l'ensemble des unités lexicales; (2) l'appariement est séquentiel, continu et immédiat; (3) bien que le réseau ne puisse être caractérisé a priori en termes d'une procédure explicite d'alignement, les observations suggèrent que les membres des cohortes, c'est-à-dire les candidats alignés à partir de la frontière de mot précédente jouent un rôle prédominant. Une augmentation brusque de l'activation de l'unité cible est observée au point d'unicité; une forte dépendance apparaît entre le point d'isolation et le point d'unicité, et les cohortes rendent compte d'une proportion importante des cas où un concurrent est activé de manière substantielle.

De ce point de vue le comportement du réseau est similaire à celui de TRACE: bien que le principe de fonctionnement suppose un appariement exhaustif, la réception séquentielle de l'information introduit une contrainte dynamique supplémentaire. Celle-ci a pour effet de restreindre les hypothèses actives de sorte que les compétiteurs de la cohorte apparaissent particulièrement influents (Frauenfelder et Peters, 1993).

Le mécanisme par lequel l'appariement est réalisé diffère cependant. Dans TRACE, l'activation des unités lexicales dépend à la fois de l'information sensorielle et des influences inhibitrices des autres candidats. De ce fait, les candidats activés précocement ont un avantage initial qui explique leur suprématie.

Par contre le réseau récurrent ne comporte pas de mécanisme de compétition directe entre les unités lexicales. L'activation des unités lexicales est entièrement déterminée par les associations entre la *trace* construite par les unités-mémoire, et les différentes cibles possibles.

Tous les candidats compatibles avec un ou plusieurs éléments de la séquence ne sont pas activés, et il n'y a donc pas appariement exhaustif. Par exemple, les rimes n'apparaissent que rarement. Cette observation implique que la trace élaborée contient plus d'information que la

séquence de segments présentée et incorpore les indices de segmentation lexicale inférés par le réseau.

Il faut cependant se garder de généraliser trop rapidement cette observation. En effet la nature des lexiques utilisés dans les quatre simulations est telle que les frontières ne sont pas fréquemment indéterminées. Des études complémentaires sont nécessaires pour déterminer dans quelle mesure le comportement global du réseau pourrait différer si le corpus présente plus d'ambiguïté.

Les performances du réseau ne sont pas limitées aux séquences particulières qui ont été utilisées durant l'apprentissage. A la fois pour la tâche C, selon la nature du prédécesseur, et pour la tâche P, selon le successeur, les résultats obtenus indiquent des performances équivalentes sur les séquences d'entraînement et sur les séquences de test. Il serait donc inapproprié de conclure que le réseau mémorise simplement des suites de mots. Ce résultat encourageant indique que le fonctionnement du réseau est relativement robuste et résiste à des variations de contexte. On peut donc penser que l'introduction progressive de nouveaux mots, qui constituerait un régime d'apprentissage plus plausible, ne devrait pas perturber massivement les connaissances acquises.

Dans une discussion publiée récemment, qui porte sur le statut psychologique des débuts de mots, Tabossi (1993) pose la question suivante: "Si, cependant, le processeur opère séquentiellement, et reçoit l'information acoustique-phonétique relative à un nouveau mot alors qu'il est encore engagé dans la 'sélection' du précédent, comment peut-il identifier cette information en temps réel comme un début de mot, et l'utiliser pour construire la cohorte qui est si cruciale pour reconnaître le nouveau mot? (p.289)".

L'architecture de réseau que nous avons explorée pourrait fournir un solution originale à ce problème, en utilisant les connaissances lexicales pour construire une *trace structurée distribuée*. La double tâche que nous avons utilisée permet au système d'extraire une information sur le mot courant, qui pourrait être utilisée par les autres composantes du processeur de langage si elle est suffisamment claire. Lorsqu'elle ne l'est pas, il suffit d'attendre. Mais l'incertitude sur la frontière n'empêche pas l'isolation rapide du mot suivant.

RÉFÉRENCES

Abu-Bakar, M. (1993). *Processing speech at varying rates: a connectionist approach.* Paper presented at the International Conference on the Psychology of Language and Communication, Glasgow, August 1993.

Bard, E. (1990). Competition, Lateral Inhibition, and Frequency. In G.T.M. Altmann (Ed.) *Cognitive Models of Speech Processing.* Cambridge, Mass.: M.I.T. Press, 185-210.

Church, K. (1987). Phonological parsing and lexical retrieval. *Cognition*, **25**, 53-69.

Elman, J.L. (1990) Finding structure in time. *Cognitive Science*, **14**, 179-211.

McClelland, J.L et Elman, J.L. (1986) The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

Frauenfelder, U.H. (1991). Une introduction aux modèles de reconnaissance des mots parlés. In Kolinsky, R., Morais, J. et Segui, J. (Eds.), *La reconnaissance des mots dans les différentes modalités sensorielles: études de psycholinguistique cognitive.* Paris: P.U.F., 7-36.

Frauenfelder, U.H. et Peeters, G. (1990). On lexical segmentation in TRACE : an exercise in simulation. In In G.T.M. Altmann (Ed.), *Cognitive Models of Speech Processing.* Cambridge, Mass.: M.I.T. Press, 50-86.

Frauenfelder, U.H. et Peeters, G. (1993). *Simulating the time-course of word recognition: an analysis of lexical competition in TRACE.* Manuscrit non publié.

Frauenfelder, U.H. et Tyler, L.K. (1987). The process of spoken word recognition: an introduction. *Cognition*, **25**, 1-20.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm.*Perception and Psychophysics*, **28**, 267-283.

Grosjean, F. (1985).The recognition of words after their acoustic offset : evidence and implications. *Perception and Psychophysics*, **38**, 299-310.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W. (1990). Activation, competition and frequency in lexical access. In G.T.M. Altmann (Ed.),*Cognitive Models of Speech Processing.* Cambridge, Mass.: M.I.T. Press, 148-172.

Marslen-Wilson,W. et Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Norris, D. (1990). A Dynamic-net model of human speech recognition. In Altmann, G.T.M. (Ed.) *Cognitive Models of Speech Processing.* Cambridge, Mass.: M.I.T. Press, 87-104.

Rumelhart, D.E., Hinton, G.E. et Williams, R.J. (1986). Learning internal representations by error propagation. In D. Rumelhart, J.L. McClelland and the PDP research group, *Parallel Distributed Processing : Explorations in the Microstructure of Cognition.*, vol. 1. Cambridge, Mass.: M.I.T. Press.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In Altmann, G.T.M. (Ed.) *Cognitive Models of Speech Processing.* Cambridge, Mass.: M.I.T. Press, 24-49.

St John, M. et McClelland, J. L. (1990). Leaning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, **46**, 217-257.

Tabossi, P. (1993). Connections, competitions and cohorts : comments on the chapters by Marslen-Wilson; Norris; and Bard and Shillcock. In G.T.M. Altmann et R. Shillcock (Eds.), *Cognitive models of speech processing. The second Sperlonga meeting.* Hillsdale: Erlbaum, 277-294.

LISTE ALPHABETIQUE DES AUTEURS

NOMS	Pages
ABRY C.	421
ADJOUDANI A.	143 et 147
AIT-MESSAOUD H	315
ALISSALI M.	125
ANDRE-OBRECHT R.	165, 169 et 175
ANGOLA O.	147
ARTIERES T.	275
ATAL B.	467
AUBERGE V.	315 et 373
AUTESSERRE D.	287
AZAMI Z.	235
BACRI N.	505
BANEL M.H.	505
BARRAS C.	491
BARTKOVA K.	181
BAUDRY M.	175
BEAUTEMPS D.	119
BECHET F.	479
BELRHALI R.	297, 303 et 373
BENE J.M.	385
BENOIT C.	143 et 147
BERGER-VACHON C.	203
BIMBOT F.	193 et 467
BOE L.J.	41, 297, 315 et 373
BONASTRE J.F.	259 et 321
BONNEAU A.	433
BOUABANA S.	439
BOUCHNAFI A.	215
BOUDELAA S.	355
BOUDRAA B.	345
BOUDRAA M.	345
BOULA de MAREUIL P.	95
CAELEN J.	131
CANDILLE L.	485
CARATY M.J.	255 et 491
CARRE R.	485
CARTIER M.	31
CASTELLI E.	107
CHOLLET G.	269
CLEMENT J.	57
COHEN M.	147
CONTENT A.	535
COSTE-MARQUIS S.	473
COURTIN J.	297
DAHAN D.	63
DAUTREMONT M.	159
DE CALMES M.	309
DE MORI R.	11
de TOURNEMIRE S.	75
DELAIS E.	379
DELEMAR O.	497
DEPPE M.	265
DESHAIES D.	81
DI BENEDETTO M.G.	339
DIANA A.	401
DJEZZAR L.	433
DUBREUIL C.	385
DUCLOS J.C.	385
DUJARDIN D.	297
DUTOIT T.	69
ENDRES W.	265
ESKENAZI M.	281
FAGYAL Z.	511
FARHAT A.	169
FARID M.	335
FENG G.	187
FISCHER M.	265
FOHR D.	199
FRANCES A.	303
FRENKENBERGER S.	125
GABIOUD B.	119
GAGNOULET C.	31
GALLINARI P.	275
GEORGE M.	427
GERARD C.	57
GILLES P.	479
GILLOIRE A.	247
GLORENNEC P.Y.	153
GOSSART A.	159
GROSSETTI M	3
GUAITELLA I.	81 et 389
GUERIN B.	345
GUIARD-MARIGNY T.	143 et 147
HATON J.P.	199
HOMAYOUNPOUR M.	269
IGOUNET S.	479
JOSPA P.	113 et 427
JOT J.M.	247
JOUVET D.	159 et 223
KAUFFMANN I.	203
KOLINSKY R.	523
KOMMENDA M.	125
LACOUME J.F.	247
LAMBLIN C.	241

NOMS	Pages
LAPRIE Y.	209 et 433
LARREUR D.	181
LAURET B.	453
LE FLOCH J.L.	255
LE GOFF B.	143 et 147
LE MEUR P.Y.	137
LEICH H.	69
LEVIN E.	467
LIENARD J.S.	339
LOUALI N.	459
MAEDA S.	41, 409 et 439
MAILLAND A.	309
MALBOS F.	175
MALDEREZ I.	361
MARI J.F.	199
MARIN M.	247
MELONI H.	101,259, 479 et 485
MERCIER G.	153
MERCIER L.	209
METAYER I.	181
MEUNIER C.	287 et 447
MOKBEL C.	223
MONNE J.	223
MONTACIE C.	255 et 491
MORAIS J.	523
MOREAU N.	241
MOUNIER-KUHN P.E.	23
MOUNIR J.	395
MOUSTY P.	529
MUNTEANU P.	291
MURGIA C.	187
NOCERA P.	321
NORMAND J.C.	385
OKAWA S.	193
ORLIAGUET J.P.	415
OUELLET M.	367
PACHES-LEAL P.	223
PACHIAUDI G.	385
PARADIS C.	81
PARLANGEAU N.	169
PASDELOUD V.	523 et 529
PEAN V.	281
PERENNOU G.	309
PERRIER P.	41 et 415
PERRIN E.	203
PIERACCINI R.	467
PITERMANN M.	131
PLANTE F.	203
POPESCU A.	241
POTAPOVA R.K.	333
PUECH G.	459
PUEL J.B.	165
QUELAVOINE R.	321
RADEAU M.	529
RHARDISSE N.	421
ROBERT J.B	37
ROBINEAU P.	101
SAVARIAUX C.	415
SCHNABEL B.	125
SCHNEIDER J.J.	101
SCHOENTGEN J.	235
SELOUANI S.A.	345
SERIGNAT J.F.	287 et 291
SHIRAI K.	193
SOKOL R.	153
SOQUET A.	229 et 427
SPRIET T.	485
STAUDACHER-V. G.	351
STERNON P.	535
VAISSIERE J.	87 et 269
VAXELAIRE B.	325
VERNET M.	385
VESCOVI C.	107
WARUSFEL O.	247
WAUQUIER-G. S.	517
WINDHEUSER C.	193
WOLINSKI K.	153
YEOU M.	409
ZEILIGER J.	287
ZNAGUI I.	215