UNIVERSITE LIBRE DE BRUXELLES

1834 - 1984

13èmes JOURNEES D'ETUDE DU GROUPE DE LA " COMMUNICATION PARLEE "

Sous le Patronage du

Groupement des Acousticiens de Langue Française

et de

l'Association Belge des Acousticiens

[redacted]

ées par
UT de PHONETIQUE
VERSITE LIBRE de
BRUXELLES

Bruxelles
28-30mai 1984

Groupement des Acousticiens de Langue Française

13 èmes JOURNEE D'ETUDE DU GROUPE DE LA " COMMUNICATION PARLEE "

LES 13es JOURNEES D'ETUDE SUR LA PAROLE qui se sont déroulées les 28, 29 et 30 mai 1984 à l'U.L.B. ont été organisées par l'Institut de Phonétique sous le patronage du GROUPEMENT DES ACOUSTICIENS DE LANGUE FRANCAISE et de l'ASSOCIATION BELGE DES ACOUSTICIENS.
Ce volume contient le résumé des communications orales et affichées et le texte complet des exposés présentés lors du thème consacré à "L'information linguistique contenue dans le signal acoustique : analyse et invariance".

M. WAJSKOP.

Réalisé avec l'aide financière du Ministère de l'Education Nationale.

Nous tenons à exprimer nos plus vifs remerciements à Monsieur le Ministre de l'Education Nationale, représenté à ces Journées par Monsieur A. BODSON, Chef de Cabinet ainsi qu'à Monsieur le Professeur H. HASQUIN, Recteur de l'UNIVERSITE LIBRE DE BRUXELLES pour l'aide matérielle qu'ils nous ont fournie.

Pour le Groupe de la Communication Parlée

Le Président, L.J. BOË

TABLE DES MATIERES

RECONNAISSANCE - Président: J.P. HATON (Université de Nancy).

COMMUNICATIONS AFFICHEES : PRODUCTION ET SYNTHESE

ANALYSE ET TRANSMISSION

RECONNAISSANCE

PERCEPTION

ANALYSE ET PRODUCTION - Président: R. DESCOUT (C.N.E.T.)

ORGANISATION D'UN SYSTEME DE GESTION DU DIALOGUE ORAL HOMME-MACHINE

D. BEROULE : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - FRANCE

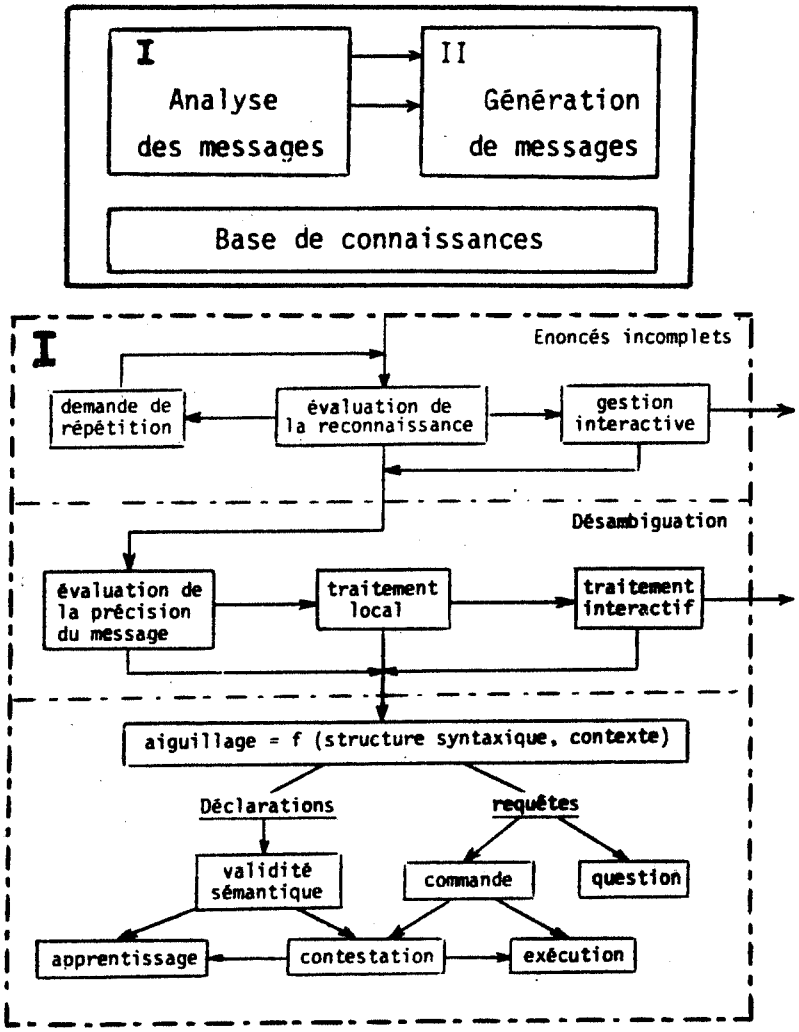
Le Traitement du dialogue oral homme-machine se différencie de la Compréhension du langage écrit. L'indéterminisme du message vocal se situe en effet à tous les niveaux, la variabilité du signal acoustique s'ajoutant à la diversité des structures admises dans le langage parlé. Heureusement, le caractère interactif du dialogue doit permettre de pallier cet indéterminisme en donnant au système la possibilité d'interroger son utilisateur pour compléter un énoncé mal reconnu ou préciser un énoncé ambigu.

Le système présenté possède des niveaux de prétraitement, éventuellement interactif, dont le rôle est d'améliorer la précision du message au niveau acoustique puis aux niveaux supérieurs. La stratégie de traitement employée ne comporte pas de prédiction sémantico-pragmatique, ce qui permet l'existence d'un module de contestation et favorise d'autre part l'apprentissage de notions nouvelles.

Deux prototypes admettant cette organisation ont été réalisés, en liaison avec une carte de reconnaissance globale et une carte de synthèse à partir du texte. L'un d'eux permet de communiquer avec un robot manipulant des objets élémentaires. Les interactions de ce logiciel avec son environnement sont les suivantes :

- Gérer les échanges d'information avec les unités d'entrée et de sortie vocales, suivant un protocole pré-établi.
- Traduire les assertions du locuteur, exprimées en langage pseudo-naturel, pour les transmettre sous forme symbolique non-ambiguë au système robotisé.
- Interpréter les messages formels envoyés par le système robotisé.

Nous poursuivons par ailleurs une recherche plus fondamentale ayant pour objet la réalisation d'un modèle cognitif aux retombées possibles dans le domaine du dialogue homme-machine.


I
Analyse
des messages
II
Génération
de messages
Base de connaissances
I
Enoncés incomplets
demande de
répétition
évaluation de
la reconnaissance
gestion
interactive
Désambiguation
évaluation de
la précision
du message
traitement
local
traitement
interactif
aiguillage = f (structure syntaxique, contexte)
Déclarations
requêtes
validité
sémantique
commande
question
apprentissage
contestation
exécution

EXPERIENCES SUR UN VOCODEUR A CLASSIFICATION POUR LA TRANSMISSION DE LA PAROLE A TRES FAIBLE DEBIT

L. MICLET, M. DABOUZ : ENST, Département Systèmes et Communications,
ERA CNRS 1044 - 46, rue Barrault - 75634 PARIS CEDEX 13.

Les études ont porté sur la conception et la réalisation d'un vocodeur à classification, c'est-à-dire d'un système de codage vectoriel spectral de la parole. Le vecteur spectral est calculé toutes les 10 ms, sur une fenêtre de 20 ms, par prédiction linéaire ; l'énergie et le pitch sont transmis séparément. Les points de l'étude ont principalement été :

- Choix d'un espace de représentation spectral, et choix de la métrique associée.
- Comparaison d'algorithmes d'apprentissage pour le dictionnaire des prototypes.
- Etude d'une méthode de décision rapide pour la recherche du plus proche voisin dans le dictionnaire.

Les résultats ont fourni une parole de qualité "comparable" au codage LPC 2400, pour un débit trois fois inférieur, et sans optimisation sur le codage de l'énergie et du pitch, ni de compression temporelle supplémentaire. Les résultats originaux portent sur :

1 - L'utilisation comparée de deux distances spectrales, l'une très simple (distance de Chebichev sur l'autocorrélation du signal), l'autre plus complexe mais réputée bonne du point de vue acoustique (distance euclidienne cepstrale). La conclusion principale est que la première distance s'avère de qualité suffisante moyennant un filtrage du signal, et une augmentation raisonnable de la taille du dictionnaire.

2 - La comparaison d'algorithmes d'apprentissage automatique du dictionnaire. Là encore, les résultats plaident en faveur de la simplicité et de la rapidité : il n'y a pas d'avantage marquant à utiliser un algorithme complexe (Lyod-Gray, ou Nuées dynamiques).

3 - La technique de décision. Il s'est avéré qu'une méthode très rapide et "approximative" de recherche du plus proche voisin dans le dictionnaire n'introduisait qu'une faible erreur, négligeable en pratique par rapport à l'approximation inhérente au procédé de codage vectoriel. La décision peut donc se faire logarithmique par rapport à la taille du dictionnaire, ce qui permet d'en augmenter (éventuellement considérablement) la taille sans sortir d'un temps

de calcul "réel". Cette recherche a donné lieu à un approfondissement et une généralisation qui ont fait le sujet d'un article dans la revue "Pattern Recognition Letters".

L'ensemble du travail et la bibliographie détaillée sont dans le rapport de thèse de M. DABOUZ.

BIBLIOGRAPHIE

L. MICLET , M. DABOUZ. Approximative fast nearest neighbour recognition. Pattern Recognition Letters 1983.

M. DABOUZ. Transmission de la parole à faible débit par vocodeur à classification. Thèse de Docteur-Ingénieur ENST Janvier 1984.

TRANSCRIPTION DE STENOTYPES EN FRANCAIS ECRIT

G. ADDA, F. NEEL : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX
C. FLUHR : Université PARIS-SUD - 91405 ORSAY CEDEX
C. MOREL : Centre Commun d'Etudes de Télédiffusion et Télécommunications
35000 RENNES

Dans cette étude, première étape vers la dictée vocale, notre approche a été de choisir en remplacement du code phonétique, le code sténotypique, ceux-ci étant très voisins. Les niveaux d'ambiguïté de ces deux codes sont en effet comparables : d'une part, un système de reconnaissance fournit un treillis phonétique, d'autre part la chaîne sténotypique est, elle, intrinsèquement ambiguë ; par exemple, le mot sténotypique 'SEL' peut se traduire par zèle, selle, serre, sel, etc ... De plus,la sténotypie fournissant un découpage du langage parlé en syllabes, il est nécessaire, comme dans le cas phonétique, de retrouver le découpage réel en mots.

L'outil de transcription proprement dit est fondé sur une analyse syntaxique qui comprend plusieurs étapes de sélections :

- La première est une sélection lexicale, qui permet de déterminer toutes les phrases constituées de mots du français, et correspondant à la chaîne sténotypique donnée. Pour cela, nous disposons d'un dictionnaire de 110.000 formes sténotypiques, correspondant à 270.000 formes orthographiques, auxquelles sont adjointes des informations linguistiques (catégories grammaticales, genre, nombre, temps et mode, etc...).

- La deuxième étape est une analyse syntaxique, utilisant les matrices de précédence binaire et ternaire , obtenues par apprentissage (et développées par A. ANDREEWSKY, F. DEBILI et C. FLUHR). Le principe de cette analyse est de sélectionner les phrases orthographiées candidates ne comportant que des couples, puis des triplets de valeurs grammaticales reconnues par les matrices de précédence. La matrice de précédence ternaire étant fréquentielle, cela permet de plus de classer les phrases sélectionnées.

- La troisième étape est une sélection par un critère de nombre de mots minimum dans la phrase, critère empirique, mais très sélectif et justifié par l'expérience.

- Enfin, on affine l'analyse à l'aide de règles d'accord simples, afin de présenter en fin de traitement, une chaîne correctement orthographiée.

L'application de ce système est, dans un premier temps, le sous-titrage en temps réel d'émissions télévisées. Ceci implique, entre autres, le développement

d'outils d'analyse propres au langage parlé, car celui-ci possède une syntaxe particulière.

D'autre part, nous avons l'intention d'améliorer la qualité des résultats en utilisant en particulier, des règles d'accord plus puissantes que celles existantes.

Le système est actuellement implanté sur gros ordinateur (NAS 9080), mais nous envisageons d'intégrer le système sur un mini-ordinateur, de le munir d'une entrée sténotypique réelle, et d'effectuer des expériences de sous-titrage en temps réel.

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

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

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

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

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

EXIGENCES

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

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

plus résistifs aux variations contextuelles. D'où une disparité entre les termes de l'opposition selon qu'ils ont une variation allophonique minimale ou maximale.

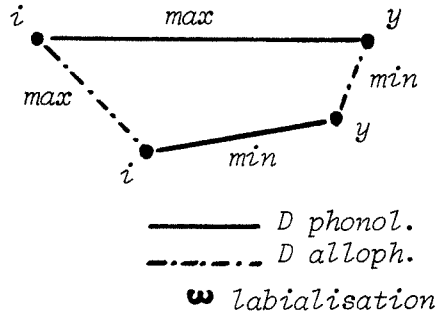
On peut donc, dans l'étude phonétique d'une opposition linguistique donnée, formuler quelques exigences sur le respect de certaines distances entre classes allophoniques (hypothèses 3 et 4). En nous limitant, pour l'exemple, à une opposition à deux termes dans deux contextes, nous définirons : D phonol., la distance entre allophones de deux phonèmes différents dans un même contexte; D alloph., la distance entre allophones d'un même phonème dans deux contextes différents.

Il faudra qu'une analyse factorielle - qui cherche, selon ses propres critères, la meilleure projection des données, de l'espace des variables sur un espace factoriel réduit - commence par respecter ces distances phonéticiennes.

LECTURES (sur des lèvres bien connues ...)

Nous disposons d'un corpus, construit depuis longtemps, permettant de tester ces hypothèses. Il contient l'opposition d'arrondissement i~y dans deux contextes s- et ʃ- (ABRY & al., 1980).

Sachant ce qu'il faut savoir de phonétique générale et française, nous nous attendions à la structure ci-contre.



Nous avons pratiqué sur nos données trois méthodes d'analyse parmi les plus connues (BENOIT, 1983) : les composantes principales normées (ANAFAC); les correspondances (ANACOR); l'analyse discriminante (ANADIS).

Nous nous sommes limités à l'étude des représentations données par les plans factoriels obtenus par analyse sur toutes les variables (8).

Sur les plans les plus "chargés" de variance et de discriminance, restait à choisir une mesure de "distance" entre nos ellipses de dispersion. Nous avons opté pour un indice de dissimilarité, utilisable quel que soit le seuil de confiance choisi pour nos ellipses (d = aire de l'ellipse "union des classes" / somme des aires des ellipses de chaque classe).

Nous avons testé nos hypothèses sur deux locuteurs extrêmes (D.L. et P.C.), l'un hyperdiscriminateur, l'autre hypo.(ABRY & al., 1980, 161-180).

Les rapports des indices de dissimilarité phonologique $d(i,y) / d(i_{\omega},y_{\omega})$ et allophonique $d(i,i_{\omega}) / d(y,y_{\omega})$ nous ont servi à mesurer le respect de nos "distances" hypothétiques.

On constate (fig. 1) que seule l'analyse des correspondances (ANACOR) est au-dessus du seuil phonologique pour les deux locuteurs, alors que toutes sont au-dessus du seuil allophonique.

Il vaut la peine de remarquer que les différences entre les trois analyses se

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

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

QU'EST-CE A DIRE ?

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

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

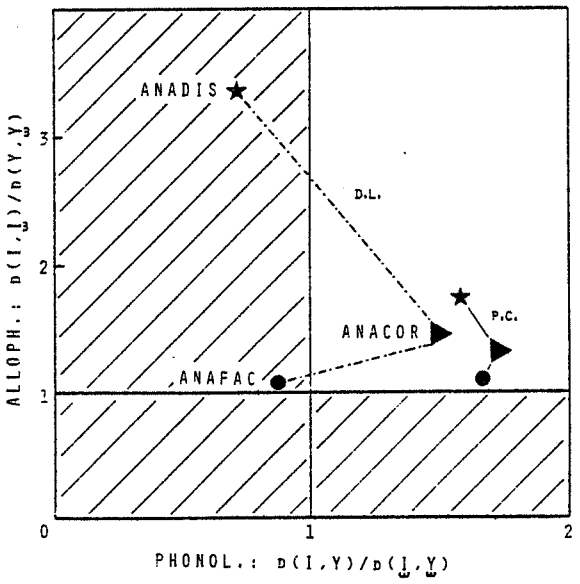


Figure 1

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

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

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

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

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

L'ACTIVITE MUSCULAIRE DU SYSTEME MANDIBULAIRE

M. GENTIL : Centre Scientifique IBM-France
Th. GAY : The University of Connecticut - Health Center, Farmington, Ct 06032, USA.

Cette étude relative au fonctionnement des muscles de la mâchoire a pour principal objectif de déterminer la spécialisation neuromusculaire en matière de parole par comparaison à d'autres fonctions mandibulaires. Au cours des quinze dernières années, un certain nombre d'études électromyographiques ont contribué à la connaissance de l'activité des muscles de la mastication (MØLLER, 1966, GAY & al., 1977, TULLER & al., 1981). Néanmoins l'organisation musculaire du système mandibulaire en ce qui concerne la production de parole est encore peu connue.

Cette expérience observait le travail de huit muscles de la mâchoire : masseters superficiel et profond, temporalis antérieur et postérieur, pterygoïde interne, les têtes inférieure et supérieure du ptérygoïde externe, ventre antérieur du digastrique. Les signaux électromyographiques étaient recueillis au moyen d'électrodes à crochets. L'enregistrement simultané des mouvements mandibulaires dans les trois dimensions (verticale, latérale et antérieure-postérieure) s'opérait grâce à un système de magnétomètres. Trois adultes mâles américains âgés d'une vingtaine d'années servaient de sujets. Ceux-ci devaient réaliser des répétitions de certains mouvements fonctionnels et produire 12 syllabes consonne-voyelle-consonne dans lesquelles les consonnes étaient [p-b-t] et les voyelles [o-i-a-u-e-æ] . Ces syllabes incluses dans une phrase porteuse étaient répétées 15 fois à deux débits de parole. Toutes les données étaient analysées par ordinateur.

Cette étude conduisait à deux constatations principales: 1)les diverses fonctions de la mâchoire quelles qu'elles soient sont basées sur des stratégies individuelles plutôt que sur des règles universelles, ce qui confirme nos précédentes observations concernant un autre articulateur, les lèvres (GENTIL & al., 1983), 2) L'espace de la mâchoire est très réduit pour la parole comparativement aux autres fonctions mandibulaires, et par conséquent les mouvements de la mâchoire sont produits avec des patterns musculaires relativement simples.

REFERENCES

GAY T., GROSS B., LIPKE D. & YAEGER J. (1977), An electromyographic study of both portions of the masseter muscle, J. Dent. Res. 56, 231 (A).

GENTIL M., GRACCO V. & ABBS J.H. (1983), Multiple muscle contribution to labial closure during speech : evidence for intermuscle motor equivalence, 11e International Congress of Acoustics, Paris.

MØLLER E. (1966), The chewing apparatus : an electromyographic study of the action of the muscles of mastication and its correlation to facial morphology, Acta Physiol. Scand. 69 (supplément 28).

TULLER B., HARRIS K.S. & GROSS B. (1981), An electromyographic study of the jaw muscles during speech, J. of Phonetics 9, pp. 175-188.

DE LA COUPE SAGITTALE A LA FONCTION D'AIRE DU CONDUIT VOCAL

H. SANCHEZ & L.J. BOË : Institut de la Communication Parlée, L.A. au C.N.R.S. n°368 - Institut de Phonétique de Grenoble.

I - INTRODUCTION

La connaissance de la fonction d'aire du conduit vocal est indispensable pour relier disposition articulatoire et signal acoustique de la parole.

Les premiers travaux associant fonction d'aire et fonction de transfert du conduit vocal remontent à 1941 : CHIBA & KAJIYAMA, à partir de moulages, prédisent et simulent la production des voyelles du japonais.

Mais dans l'état actuel des techniques, les mesures d'aire pour de la parole continue sont pratiquement impossibles à obtenir; il faudrait avoir des mesures d'aire toutes les 5 ms pour prétendre à une bonne précision.

Aussi se sont développées des procédures de mesure indirectes, citons par exemple : SCHROEDER (1967), SONDHI & GOPINATH (1971), DESCOUT & al. (1977).

Le travail que nous présentons a été effectué à partir d'un moulage du conduit vocal réalisé sur un cadavre. Disposant de mesures d'aire et de dimensions transversales, il s'est agit de déterminer le modèle optimal permettant le passage d'une dimension latérale à l'aire frontale.

A partir du moulage intial, on a obtenu plusieurs exemplaires en résine, l'un d'entre eux a été découpé sagittalement et un autre transversalement en 10 sections.

II - MODELE DE PASSAGE "COUPE SAGITTALE - FONCTION D'AIRE"

II - 1. Modélisation pharyngale et buccale

Les dimensions latérales et les aires de chaque section ont été mesurées grâce à un ensemble tablette d'acquisition*/logiciel graphique développé pour traiter les tracés acquis sous forme de contour fermé.

Pour le jeu des dimensions transversales, B_i du conduit vocal j nous avons calculé les paramètres c_j, k_i et p qui permettent une optimisation de la mesure de l'aire A_{ij} : $A_{ij} = c_j (k_i B_{ij})^p$ ... (I) avec les constantes : C_j de proportionnalité, K_i de normalisation et p d'exponentiation.

A partir de ce résultat il sera aisé de présenter un conduit vocal estimé et simplifié sous forme de sections elliptiques.

---

*. Visu 4012 et tablette 4953 TEKTRONIX. Nous remercions J. GENIN pour ses précieux conseils et J. RAYMOND pour son expérience graphique.

II - 2. Calcul des aires transversales du conduit vocal

Tout d'abord nous avons cherché pour les 10 sections dont nous disposions les paramètres c_j, k_i et p qui optimisent au mieux. En fait on regroupe les sections en zones pour obtenir le minimum de jeux de paramètres pour un maximum de précision (v.eq. (I)).

Dans un deuxième temps nous avons calculé les dimensions CD_i des ellipses ayant les mêmes aires (estimées), $A_i = 0.25\pi\ B_i\ CD_i$ ... (II)

III - RESULTATS

Après de nombreux essais, nous avons (sans tenir compte des lèvres) retenu 5 zones : le larynx (glotte), la partie laryngienne du pharynx, la partie buccale du larynx, la cavité buccale (zones uvulaire, vélaire, palatale et prépalatale) et la zone alvéolaire.

Les valeurs obtenues sont données au tableau I.

Avec p = 1.5 nous pouvons comparer nos résultats à ceux de MEMERLSTEIN (1973) avec p = 0.666 à ceux de LEFEVRE & al. (1983), pour k_i = 1 . Dans l'ensemble ils sont qualitativement similaires, il faut bien préciser que nous sommes partis de mesures effectives.

Le tableau II nous permet de comparer les erreurs obtenues par section avec deux modes différents de regroupement des sections.

Section	c	k	p	%Erreur	Mode 1 %Erreur	Mode 2 %Erreur
1	1.472	1.0	1.5	0.00	0.00	0.00
2	3.043	1.0	1.5	20.00	20.00	7.27
3	3.043	1.0	1.5	20.00	20.00	39.11
4	2.207	1.0	1.5	0.54	0.54	15.47
5	2.207	1.0	1.5	0.54	0.54	16.37
6	3.007	1.0	1.5	8.25	8.25	8.25
7	3.007	1.0	1.5	1.61	1.61	1.61
8	3.007	1.0	1.5	8.67	8.67	8.67
9	3.007	1.0	1.5	1.18	1.18	1.18
10	1.782	1.0	1.5	0.00	0.00	0.00

Tableau I - Coefficients obtenus pour l'estimation de l'aire correspondante et erreurs.

Tableau II - Erreurs obtenus par section avec les regroupements suivants :
Mode 1 : 1; 2 et 3; 4 et 5; 6, 7 8 et 9; 10.
Mode 2 : 1; 2, 3, 4 et 5; 6, 7, 8 et 9; 10.

Les figures 1, 2 permettent de visualiser les variations des aires estimées selon les regroupements choisis et présentent les erreurs.

IV - CONCLUSION

Tout en précisant bien les limites d'un tel travail : moulage unique, sur un cadavre, la méthode utilisée nous a permis d'obtenir des résultats qui,

faute de mieux, sont tout à fait utilisables dans le cadre d'une synthèse articulatoire. Cette procédure est actuellement testée par analyse-synthèse sur des radiocinématographies de production de parole.

BIBLIOGRAPHIE

CHIBA T. & KAJIYAMA M. (1941),
The Vowel : its nature and structure. - Tokyo-Kaiseisan Publishing Company, Ltd., Tokyo.

DESCOUT R., TOUSSIGNANT B., LEFEVRE J.P. & LECOURS M. (1977),
Détermination de la fonction d'aire du conduit vocal : Quantification et interpolation. - Articulatory Modeling and Phonetics; G.A.L.F., p. 31-40.

FANT G. (1960),
Acoustic theory of speech production. - S-Cravenhage : Mouton & Co.

LEFEVRE J.P., LONCHAMP F. & ZERLING J.P. (1983),
Détermination des fonctions d'aire vocaliques d'un sujet par deux méthodes. - 11e I.C.A., Paris - Lyon - Toulouse.

MERMELSTEIN P. (1973),
Articulatory motion for the study of speech production. - J.A.S.A. 53, n° 4, p. 1070-1082.

SCHROEDER M.R. (1967),
Determination of the geometry of the human vocal tract by acoustic measurements. - J.A.S.A. 41, p. 1002-1010.

SONDHI M.M. & GOPINATH B. (1971),
Determination of the vocal-tract shape from impulse response at the lips. - J.A.S.A. 49, p. 1869-1873.

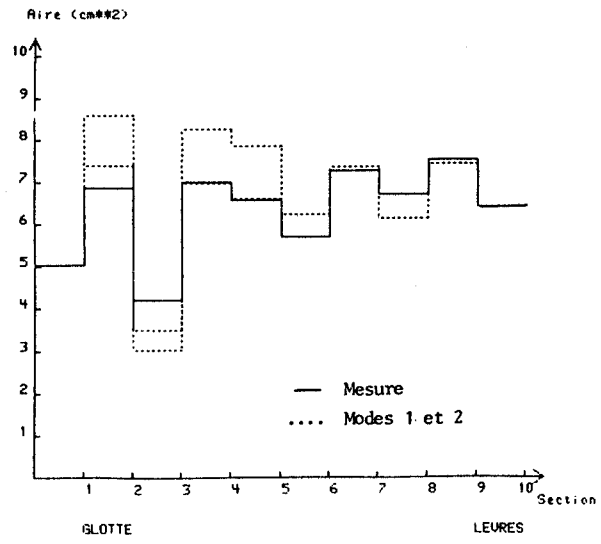


Figure 1 - Aire.

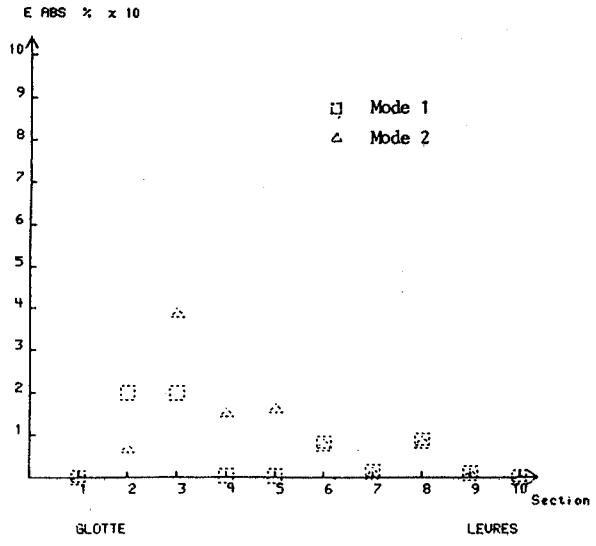


Figure 2 - % d'erreurs.

FONCTIONS DE SENSIBILITE D'UN MODELE DISSIPATIF DU CONDUIT VOCAL

F. CHARPENTIER : Département Recherche sur la Communication Parlée
C.N.E.T. 22301 LANNION - France

On présente dans cette communication l'application de l'analyse de la sensibilité des circuits électriques par *méthode lagrangienne* à l'analogue électrique du conduit vocal (F.Charpentier, "Computation of vocal-tract sensitivity functions: a lagrangian approach", ASA meeting, 1982, Orlando). Cette méthode permet de traiter des cas plus généraux que la méthode proposée par Fant et Pauli ("Spatial characteristics of vocal-tract resonance modes", Speech Communication Seminar, 1974, Stockolm), qui est fondée sur l'étude des petites variations des fréquences de résonance d'un circuit entièrement réactif. On mentionne à ce propos une démonstration élégante à l'aide du théorème de Tellegen, qui montre bien qu'on ne peut pas suivre le même raisonnement dans les cas dissipatifs correspondant à des modélisations plus réalistes du conduit vocal. Il faut alors recourir à une approche différente, consistant à étudier plutôt les petites variations de la fonction de transfert elle-même. On introduit la notion de *fonction de transfert partielle* reliant une source acoustique située à une position quelconque du conduit à la réponse acoustique en tout autre point. La méthode lagrangienne permet alors d'exprimer la sensibilité de n'importe quelle fonction de transfert partielle par rapport aux variations de n'importe quel paramètre géométrique, comme le produit, au signe près, de deux autres fonctions de transfert partielles faisant intervenir le point où se produit la variation géométrique. Cette formule est attrayante parce qu'elle permet d'exprimer analytiquement les dérivées de tous ordres des quantités acoustiques (débit volumique et pression) par rapport aux paramètres géométriques.

On présente des résultats de cette analyse appliquée à la fonction de transfert globale, qui relie le débit de la source glottique au débit volumique labial. La sensibilité de la fonction de transfert par rapport à un paramètre géométrique d'un type donné (soit aire, soit longueur) peut s'exprimer sous la forme d'une fonction scalaire de deux variables, la fréquence et l'abscisse dans le conduit, et on

peut naturellement la représenter en perspective. Sur une telle représentation, on observe des zones de sensibilité maximale autour des fréquences de résonance. De cette fonction de sensibilité à deux variables on déduit les fonctions habituelles représentant la sensibilité des fréquences formantiques. On compare les résultats obtenus par la méthode lagrangienne à ceux obtenus par la méthode de Fant. On discute l'influence des pertes de différentes natures (visqueuses, thermiques, vibratoires et celles dues au rayonnement labial) sur l'allure des fonctions de sensibilité ainsi que l'erreur commise en appliquant la formule de Fant aux cas dissipatifs. On présente également des sensibilités d'amplitude formantique dont on observe la corrélation avec les sensibilités de fréquence formantique pour les deux premiers formants.

APPROCHE ANALYTIQUE DU RAYONNEMENT DES CONDUITS ORAL ET NASAL

B. MERLIER & B. GUERIN : Institut de la Communication Parlée. CNRS LA N°368
23, rue des Martyrs - 38031 GRENOBLE CEDEX

Dans les études réalisées jusqu'à ce jour, les effets d'interaction entre le nez et la bouche, lors de l'émission de la parole, ne semblent pas avoir été envisagés : on se contentait généralement d'additionner les pressions rayonnées au nez et à la bouche.

Cette étude a pour but de préciser les caractéristiques des deux rayonnements et d'introduire un couplage entre ces deux sources, afin d'en évaluer l'importance.

APPROCHE ANALYTIQUE.

L'interaction mutuelle des deux sources vocales est un phénomène très compliqué ; c'est pourquoi celles-ci ont été modélisées par deux pistons rigides circulaires, situés dans un même plan infini. Une étude théorique préliminaire montre que **chacune des deux sources se trouve perturbée par le rayonnement de l'autre source.**

Ce fait se trouve formulé dans l'expression des pressions :

à la bouche : $p_o = Zoo.Uo + Zon.Un$,

au nez : $p_n = Zno.Uo + Znn.Un$,

où p_o, p_n sont les pressions orale et nasale,

Uo, Un les débits correspondants,

Zoo, Znn les impédances propres (impédance que présenterait chaque source **seule** dans l'espace considéré),

Zon = Zno les termes d'impédance mutuelle.

Les diverses impédances sont calculées par la formule de RAYLEIGH (PRITCHARD, 1960). Cependant, la limitation en fréquence (f<3000 Hz) de cette étude permet d'utiliser des formules approchées :

$$Z_{roo} = S_o \rho c \left[2\pi \left(\frac{f}{c}\right) S_o + j \frac{16\, f \sqrt{S_o}}{3c\sqrt{\pi}} \right]$$

Z_{rnn} : idem à l'indice près

$$Z_{ron} = \rho c \frac{S_o S_n}{cd} \left[\sin\left(\frac{2\pi fd}{c}\right) + j \cos\left(\frac{2\pi fd}{c}\right) \right]$$

où ρc caractérise le milieu de propagation, c est la vitesse du son, f la fréquence, d la distance séparant les deux sources orale et nasale, et So et Sn leur surface.

Note : La modélisation des sources vocales par deux pistons à la surface d'une sphère a été envisagée ; mais celle-ci apporte peu de modification en contre-partie du net accroissement de la complexité des calculs.

PREMIERS RESULTATS.

Les premiers résultas obtenus à l'aide d'un modèle de simulation (B.LHERM, B.GUERIN) apportent une meilleure connaissance des rayonnements aux nez et à la bouche. Le faible rayonnement au nez induit peu sur le rayonnement à la bouche, alors que le phénomène inverse est plus marqué.

Globalement, les effets du couplage restent faibles.

BIBLIOGRAPHIE :

PRITCHARD R.L. (1960), JASA **32**, pp. 730-737.

LHERM B. & GUERIN B. (1984), 13° JEP BRUXELLES.

SIMULATION HARMONIQUE ET TEMPORELLE DES CONDUITS COMPLEXES DE L'APPAREIL PHONATOIRE. APPLICATION AUX VOYELLES NASALES

B. LHERM & B. GUERIN : Institut de la Communication Parlée LA N°368
23 , rue des Martyrs - 38031 GRENOBLE CEDEX

L'originalité du modèle de production mis au point réside essentiellement dans la capacité de définir rapidement et simplement une configuration vocalique particulière, sous forme d'un assemblage arborescent de tubes. Les tubes sont connus par leur longueur, aire et facteur de forme ; les configurations possibles comportent des dérivations, des cavités, des orifices de rayonnement, en nombre quelconque, suivant le type de production de parole étudié.

Le modèle comporte deux parties qui utilisent la même définition de conduit vocal mais fournissent des résultats de types différents :

Sur la base de l'analogie classique entre une ligne acoustique et une ligne électrique artificielle, on étudie les caractéristiques harmoniques du conduit vocal. Le calcul fournit la fonction de transfert du conduit étudié (rapport de la pression recueillie aux orifices du conduit sur le débit à l'entrée, valeurs amplitude/fréquence et courbe), les pôles et les zéros (fréquences, bandes passantes et amplitudes) et les impédances d'entrée des différents tubes utilisés.

Le synthétiseur proprement dit constitue la deuxième partie du modèle. On utilise le modèle de propagation dû à KELLY & LOCHBAUM pour propager les ondes de débit glottique depuis la source jusqu'aux orifices de rayonnement. Le modèle de la source utilisé est du type "modèle à deux masses" (GUERIN B., 1978), et on calcule la pression rayonnée pour un auditeur situé à environ 30 cm du locuteur. Les paramètres d'entrée du synthétiseur sont donc, P_s et Q, la pression subglottique et un coefficient représentant la raideur des muscles du larynx, et la description de la configuration vocalique (fonction d'aire).

Dans un premier temps, ce modèle a été utilisé pour l'étude des voyelles nasales du français, qui sont produites à partir de configurations vocaliques assez complexes (couplage avec les fosses nasales, sinus, sources de rayonnement supplémentaires).

L'influence de divers paramètres a été mise en évidence, tels que l'abaissement du voile du palais, le dédoublement des fosses nasales ou la présence des cavités débouchant dans les fosses nasales. D'autres études sont menées actuellement sur le rayonnement des divers orifices de sortie de l'appareil phonatoire (B. MERLIER et al.).

BIBLIOGRAPHIE SOMMAIRE

GUERIN B. (1978)
Thèse Doct. d'Etat, I.N.P. Grenoble

KELLY J. R. Jr, LOCHBAUM C. (1962)
Proc. Stockholm-Speech Communications Seminar - R.I.T, Stockholm, p 127-130

B. LHERM (1984)
Thèse Doct. Ing., I.N.P. Grenoble

HIERARCHISATION DES INDICES PAR ANALYSE STATISTIQUE - MONO ET MULTILOCUTEUR

G. CAELEN-HAUMONT, J. CAELEN

Laboratoire C.E.R.F.I.A., Université P. SABATIER, TOULOUSE - France

INTRODUCTION

Dans le domaine de la phonétique acoustique, l'inadéquation d'un modèle d'analyse en traits binaires, a conduit les chercheurs à développer des méthodes d'investigation qui tiennent compte du graduel et du quantitatif.

Dans cette optique, le but de cet article est d'une part, de rechercher des invariants, et d'autre part, d'évaluer et de hiérarchiser 5 indices spectraux non formantiques, décrits (1) et utilisés antérieurement pour la discrimination des phonèmes et des locuteurs (2), en fonction de leur contribution statistique. Nous envisagerons simultanément les deux domaines conjoints de l'analyse interphonémique (intralocuteur) et interlocuteur (intraphonémique), en nous limitant à quelques exemples significatifs illustrant les directions de l'étude que nous menons par ailleurs.

1 CORPUS ET EXPERIMENTATION

Pour cette étude qui continue les précédentes (1)(2), nous avons utilisé la même base de données, implémentée sur MINC-DECLAB 23: celle-ci est constituée de fichiers statistiques regroupant pour chacun des 5 locuteurs masculins, 27 logatomes de structure CVCVCV, constitués comme suit:

$$\text{consonne voisée (sauf /z/, /ʒ/)} + /a/ + \begin{cases} /m/ \\ /n/ \\ /l/ \end{cases} + \begin{cases} /i/ \\ /y/ \\ /u/ \end{cases} + \begin{cases} /l/ \\ /m/ \\ /n/ \end{cases} + /a/.$$

Les phonèmes étudiés sont ceux de la syllabe médiane (/l/, /m/,/n/, /i/, /y/, /u/) et la voyelle finale /a/.

Ces fichiers sont organisés pour avoir accès et traiter les informations prélevées sur chacun des logatomes par locuteur, informations de type phrastique, prosodique, phonétique et acoustique.

2. ETUDE DES CANAUX

2.1. CORRELATIONS DES CANAUX

Utilisant les résultats de l'analyse en composantes principales, nous avons tout d'abord étudié la matrice des coefficients de corrélation des valeurs spectrales prélevées dans les 24 canaux du modèle d'oreille (3). Cette matrice a un double intérêt: elle précise d'une part les corrélations entre canaux et permet d'autre part, de procéder à la reconnaissance par décision bayésienne (4). A travers elle, il est possible d'examiner une certaine invariance phonémique et/ou locuteur.

Pour simplifier la description, la matrice numérique est convertie en matrice visuelle carrée, à l'aide de 4 degrés de couleur: blanc, gris clair, gris moyen, gris foncé, correspondant respectivement aux seuils $\pm$ 0.50, $\pm$ 0.75, $\pm$ 0.90, $\pm$ 0.95.

2.1.1. VARIABILITE INTERLOCUTEUR

En prenant comme exemple le phonème /i/(moyenne de toutes les observations par locuteur), cette variabilité apparaît nettement (cf figures 1 à 3 en annexe):

- chez le locuteur DD, deux zones à peu près égales d'intercorrélation des canaux (1-13, 14-24), mais d'expansion différente: selon la diagonale pour la première, en forme de carré pour la seconde.

- chez le locuteur PF, deux zones inégales (1-15, 16-24), mais d'expansion semblable (deux carrés).

- chez le locuteur RP, cinq zones dont quatre égales (1-10, 14-24) disposées en croix dans la matrice, d'expansion comparable (carré), mais de degré de corrélation différent (axe NO-SE privilégié), la cinquième zone étant centrale (10-14), d'expansion et de degré de corrélation comparables aux précédentes.

2.1.2. VARIABILITE INTERPHONEMIQUE

Les figures 4 et 5 fournissent un exemple de variabilité de la corrélation des canaux entre deux phonèmes très voisins /m/, /n/ chez un même locuteur, SC. Si pour /m/ et /n/, les zones d'intercorrelation sont comparables (1-3, 4-24), les expansions sont différentes: diagonales (3-12) et carré (13-24) pour /m/, carré (4-24) pour /n/.

2.1.3. INVARIANCE PHONETIQUE

Bien que la variabilité, dans les deux domaines, interlocuteur et interphonémique, soit manifeste, les canaux de moindre corrélation sont à peu près fixes: il existe une correspondance assez nette avec les zones formantiques. Les figures 1 à 5 témoignent de ce fait. Pour illustrer plus précisément ces propos, nous prendrons l'exemple du phonème /i/. La figure 6 représente les groupes d'intercorréations des canaux de ce phonème, pour les 5 locuteurs.

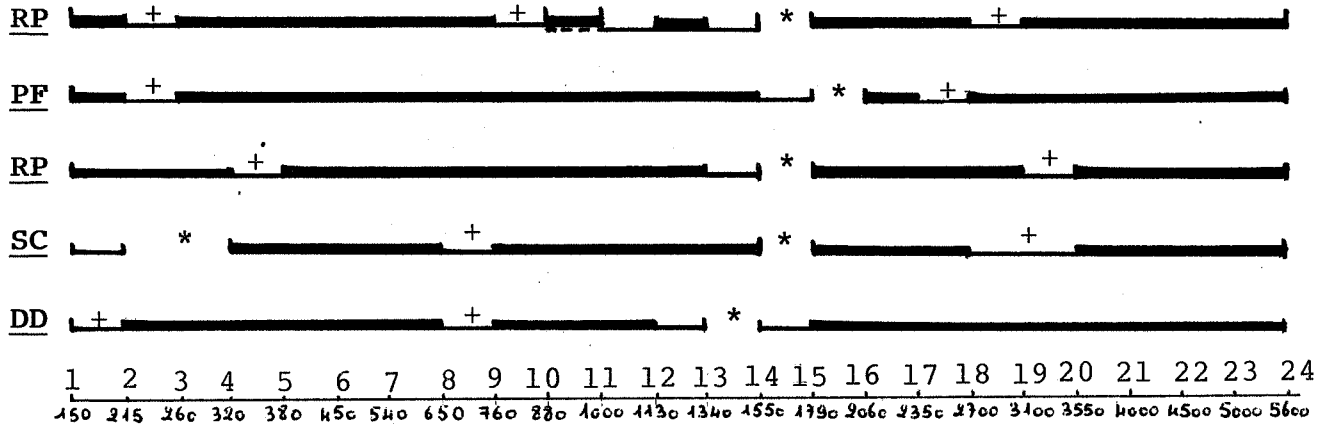


FIGURE 6: Représentation de l'intercorrélation des canaux du phonème /i/ (perspective interlocuteur)

Etudiant les cinq matrices de /i/, correspondant aux valeurs moyennes de toutes les observations par locuteur, nous avons distingué deux niveaux de rupture de corrélations des canaux, les grands ensembles (limites symbolisées par le signe "*"), leurs sous-ensembles (signe "+"). Le graphique établit par ailleurs la correspondance des canaux sur l'échelle fréquentielle.

On remarque par ordre d'importance décroissante, une zone de rupture (ou limite de corrélation) bien marquée dans la zone 1600-1800, ce qui correspond sensiblement au deuxième formant, puis une zone dans les basses fréquences, 215-320 (1er formant), puis dans la zone du 3ème formant (2700-3000), et une autre parfois attestée, aux alentours de 760 ou un peu plus, ce qui semble confirmer les observations faites par certains chercheurs (M. ROSSI et alii, 1978).

2.2. ANALYSE EN COMPOSANTES PRINCIPALES

2.2.1. REDUCTION DES PARAMETRES

En analyse en composantes principales, le pourcentage cumulé des valeurs propres donne le taux de variance expliquée. Inversement, en se donnant un seuil de 99%, on peut en déduire le nombre de combinaisons linéaires nécessaires pour décrire au mieux l'information (cf tableau 1), selon les phonèmes et les locuteurs:

	/i/	/y/	/u/	/A/	/l/	/m/	/n/
DD	5	5	5	5	6	5	5
SC	5	5	5	3	6	6	5
PP	4	4	5	6	6	5	5
PF	5	4	4	5	5	4	5
RP	4	6	5	5	5	4	4

TABLEAU 1: nombre de combinaisons linéaires nécessaires pour un seuil de 99%

Si l'on définit un indice comme une combinaison linéaire de canaux, 3 à 6 combinaisons sont nécessaires (moyenne 4.9): en moyenne, 5 indices sont donc suffisants pour décrire l'échantillon considéré (à 99%). Bien que ce nombre soit assez stable, on remarque cependant une plus grande variabilité pour le phonème ou plutôt l'archiphonème /A/ (3,5,6 combinaisons) et pour le locuteur SC (idem).

2.2.2. CONTRIBUTION DES CANAUX

L'examen de la contribution des canaux sur les axes factoriels, fournit de manière complémentaire à l'étude de la corrélation, des informations sur la dispersion des observations.

Il ressort de cette étude, l'idée d'une certaine variabilité au sein d'une parenté de structures plus ou moins affirmée. Ainsi, tableau 2, nous avons considéré par exemple, deux phonèmes proches /m/ et /n/, pris selon les deux perspectives d'analyse interphonémique (opposition /m/-/n/, axe vertical) et interlocuteur (opposition SC-PP pour /n/, axe horizontal).

Dans les deux cas, nous avons relevé sur les trois premiers axes factoriels (I,II, III), le numéro des canaux, qui au sein de groupes rangés par ordre de contribution décroissante (au maximum 3 groupes donc 3 canaux principaux, numérotés 1,2,3), dominent leurs voisins.

Du point de vue intralocuteur, 3 canaux (9,3,11) sont communs aux deux phonèmes, et 7 sont spécifiques du phonème (/m/: 20,1,7; /n/: 21,14,18,4). Par ailleurs, les canaux communs ne se trouvent pas nécessairement sur le même axe factoriel, ce qui tend à montrer que les nuages de dispersion n'ont pas une configuration semblable.

Du point de vue interlocuteur, 2 canaux (13,14), sont communs aux locuteurs SC et PP pour /n/, et 9 sont spécifiques de chacun d'eux.

		AXES FACTO-RIELS	INTERLOCUTEUR : /n/					
			SC			PP		
INTERPHONEMIQUE SC	/n/	I	21	9	14	14	24	3
		II	11	18	4	19	6	/
		III	3	/	/	1	/	/
	/m/.	I	20	9	1			
		II	3	/	11			
		III	7	/	/			
			1	2	3	1	2	3
			ORDRE D'IMPORTANCE DE CONTRIBUTION DES CANAUX					

TABLEAU 2 : Contribution des canaux dans les domaines interlocuteur et interphonémique.

Discussion

Tous ces exemples montrent la difficulté de trouver des invariants en ce domaine, difficulté due, semble-t-il, à la finesse de la définition spectrale. Le nombre des canaux est en effet trop grand pour obtenir une localisation fixe et stable des informations en multilocuteur, car des zones entières du spectre sont corrélées et donc redondantes. L'intérêt des indices que nous avons définis (1), considérés comme des combinaisons linéaires de canaux, est de réduire le nombre de paramètres (dans un rapport de 5), pour un taux d'information tout-à-fait comparable (99%). Les contributions des canaux sur les axes cependant ne sont pas invariantes, ce qui interdit tout système d'indices optimal construit sur ce principe.

3. ETUDE DES INDICES

3.1. NATURE ET FONCTION

Comme nous l'avons constaté ci-dessus (cf 2.1.3.), il existe un lien entre zones formantiques et zones de corrélation spectrale minimale. Des combinaisons invariantes de canaux n'existant pas, il est donc raisonnable de s'appuyer sur les zones formantiques pour définir des indices (J. CAELEN et G. CAELEN, 1981): fermé/ouvert (FO), aigu/grave (AG), bémolisé/diésé (BD), doux/strident (DS), et écarté/compact

(EC). Le sixième indice, continu/discontinu (CD), reposant à la différence des autres, sur une variation temporelle, n'a pas été retenu dans cette étude.

Une originalité des indices est leur caractère de non-binarité et en outre, leur capacité discriminatoire des phonèmes et des locuteurs (G. CAELEN-HAUMONT et N. VIGOUROUX, 1983). Pour pallier cependant à l'inconvénient de la variabilité des combinaisons linéaires, on pourrait, dans un système phonétique descriptif, ou dans un système de reconnaissance, pondérer différemment (ou hiérarchiser) les indices entre eux. Le nombre de ces indices n'étant pas optimal, il est à craindre qu'ils soient corrélés, corrélation que la hiérarchisation doit prendre en compte.

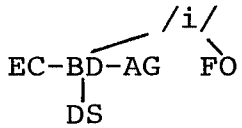
3.2. HIERARCHISATION DES INDICES

Pour plus de clarté, l'analyse qui suit, a recours à des schémas arborescents sur lesquels figurent quelques symboles. Par convention, dans les schémas, l'axe horizontal représente l'intercorrélation, et l'axe vertical la non-corrélation. L'ordre d'écriture des indices reproduit l'ordre décroissant des contributions (droite à gauche, haut vers le bas) sur chacun des axes. On retient pour l'étude les 3 premiers axes, numérotés de 1 à 3. Ces trois axes sont orthogonaux entre eux, et nous relions par le signe | les indices ayant des contributions maximales sur deux axes orthogonaux. Les schémas ou structures type des phonèmes ou de l'espace locuteur, représentent une synthèse des phénomènes observés: un indice absent dénote une grande instabilité de position.

3.2.1. PERSPECTIVE INTERLOCUTEUR

- **phonème /i/:**

.Structure type:



.Invariants:

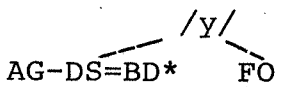
-BD,EC toujours en 1
-FO jamais en 1
-FO jamais corrélé aux 4 autres
-EC,BD toujours corrélés
-FO | EC,BD,DS

.Variabilité:

La contribution des indices AG et DS est parfois très différente selon les locuteurs: ce fait est en accord avec la variabilité de la quantité d'énergie des troisièmes et quatrièmes formants.

- **phonème /y/:**

.Structure type:



.Invariants:

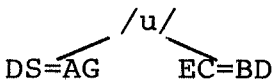
-AG toujours en 1
-FO jamais en 1
-FO | DS,AG

.Variabilité:

Elle est plus grande pour /y/ que pour /i/. Le phonème /y/ du locuteur DD est proche de /i/ du locuteur PF.

- **phonème /u/:**

.Structure type:



.Invariants:

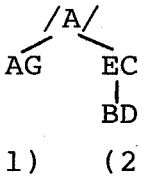
-EC toujours corrélé
-FO | DS,EC

. Variabilité:

Chez les cinq locuteurs, on constate pour /u/ des corrélations plus instables que pour les phonèmes précédents. Concernant les voyelles fermées, on remarque une certaine invariance (FO | DS).

- **phonème /a/:**

.Structure type:



.Invariants:

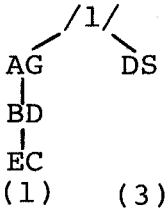
-FO jamais corrélé
-FO | BD,EC

.Variabilité:

EC et BD ont un comportement voisin; ils se trouvent toujours sur le même axe, quel qu'il soit, et sont parfois corrélés: locuteurs DD,SC. Par ailleurs, la variabilité que l'on constate à propos de la contribution des indices sur les axes est à mettre en relation avec la variabilité de réalisation de l'archiphonème /A/.

- **phonème /l/:**

.Structure type:



.Invariants:

-FO jamais corrélé
-FO | BD,AG
-EC | DS

---

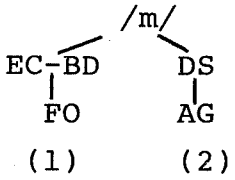
* Le signe = indique que la contribution des indices est la même.

. Variabilité

En règle générale, les indices sont le plus souvent non corrélés entre eux. On constate une grande variabilité de structure selon les locuteurs, variabilité que l'on retrouve par ailleurs, sur les spectres eux-mêmes. Cette variabilité réside dans le fait que la corrélation ne s'établit pas toujours entre les mêmes indices.

_ **phonème /m/:**

.Structure type:

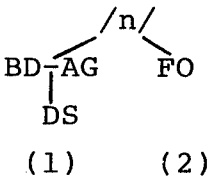


.Invariants:

-DS jamais en 1
-DS,FO jamais corrélés aux 4 autres
-FO | AG
-DS | EC,BD

- **phonème /n/:**

.Structure type:



.Invariants:

-DS,FO jamais corrélés aux 4 autres
-EC | DS
-FO | BD

.Variabilité:

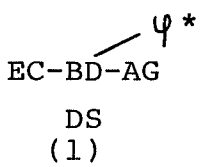
Les indices BD et EC sont moins corrélés que pour /m/, seuls les locuteurs SC et PP attestent cette corrélation.

3.2.2. PERSPECTIVE INTRALOCUTEUR

Il s'agit de définir, si possible, une structure spécifique du locuteur en fonction, non de l'ensemble des phonèmes du français, mais sur la base des phonèmes de l'étude. L'intérêt est de comparer les contributions des indices entre les 5 locuteurs relativement à un ensemble fixe de phonèmes.

- **locuteur DD:**

.Structure type:



.Invariants:

-FO jamais corrélé
-FO | AG,DS

---

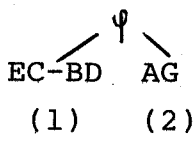
* Le signe ψ symbolise les 7 phonèmes de l'étude.

.Variabilité:

On remarque une instabilité de la contribution des indices sur les axes 2 et 3.

- **locuteur SC:**

.Structure type:



.Invariants:

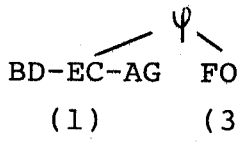
-FO jamais corrélé
-AG | EC,BD

.Variabilité:

On constate une variabilité plus importante chez le locuteur SC que chez le locuteur DD.

- **locuteur PP:**

.Structure type



.Invariants:

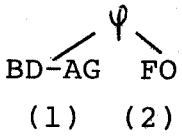
-AG toujours en 1
-FO jamais corrélé
-FO | EC,BD

.Variabilité:

BD et AG, comme pour le locuteur PF, sont souvent corrélés entre eux (phonèmes /i/,/y/,/u/,/l/).

- **locuteur PF:**

.Structure type:



.Invariants:

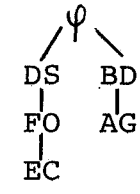
-AG toujours en 1
-DS jamais corrélé
-FO | BD

.Variabilité:

BD et AG sont très souvent corrélés entre eux (phonèmes /i/,/y/, /A/,/m/,/n/), alors que les autres indices ont des intercorrélations très peu nombreuses.

- **locuteur RP:**

.Structure type:



.Invariants:

-FO jamais corrélé
-FO | AG

.Variabilité:

Les corrélations sont peu stables. Ce locuteur se différencie très sensiblement des autres locuteurs.

D'une manière générale, on constate que les locuteurs DD, PP, PF présentent une contribution des indices sur les axes moins instable, ce qui implique des corrélations ou inversement des non-corrélations, mieux affirmées. Tous ces résultats concernant les locuteurs, sont à confirmer dans le cadre d'une étude plus vaste.

4. CONCLUSION

Cette étude nous permet de mieux définir l'interêt de nos 5 indices. En effet, par rapport aux canaux, les indices offrent:

-une concentration de l'information
-une redondance moindre
-une perte d'information négligeable (<0.1%).

De ce fait, les invariants y apparaissent plus nettement. Cependant, pour certains locuteurs et/ou phonèmes, des indices sont corrélés, essentiellement BD et EC pour les voyelles fermées, ce qui permet de les mettre au même niveau dans la plupart des arbres hiérarchiques. Inversement, FO et DS sont non seulement peu corrélés, mais se trouvent également sur des axes distincts. On définit ainsi des structures type de phonèmes et/ou de locuteurs par une méthode qui devra être utilisée sur des corpus plus vastes et plus variés. Par ailleurs, des applications en reconnaissance (niveau descendant) peuvent d'ores et déjà être envisagées.

REFERENCES BIBLIOGRAPHIQUES

(1) J. CAELEN et G. CAELEN, Indices et propriétésdans le projet ARIAL II, Actes du Séminaire Encodage et Décodage Phonétique, GALF-CNRS, TOULOUSE, 1981.

(2) G. CAELEN et N. VIGOUROUX, Les indices de distribution spectrale: étude comparative au travers de 2 analyses discriminantes monolocuteur et interlocuteur, Speech Communication 2,1983,133-136.

(3) J. CAELEN Un modèle d'oreille. Analyse de la parole continue. Reconnaissance phonémique. Thèse d'état, TOULOUSE, 1979.

(4) J. WOLF et alii, Further investigation of probabilistics methods for text-independent speaker identification, Proceedings of ICA, Speech and Signal Processing, BOSTON MA., USA, 1983.

(5) M. ROSSI, C. LE CORRE, G. MERCIER, Indices de détection de formants sur analyse spectrale par canaux, 9èmes JEP, GALF-CNRS, LANNION, 1978.

(6) C. ABRY et L-J. BOË, Essai d'anlyse phonétique des indices du voisement, 9èmes JEP, GALF-CNRS, LANNION, 1978.

figure 1
locuteur DD

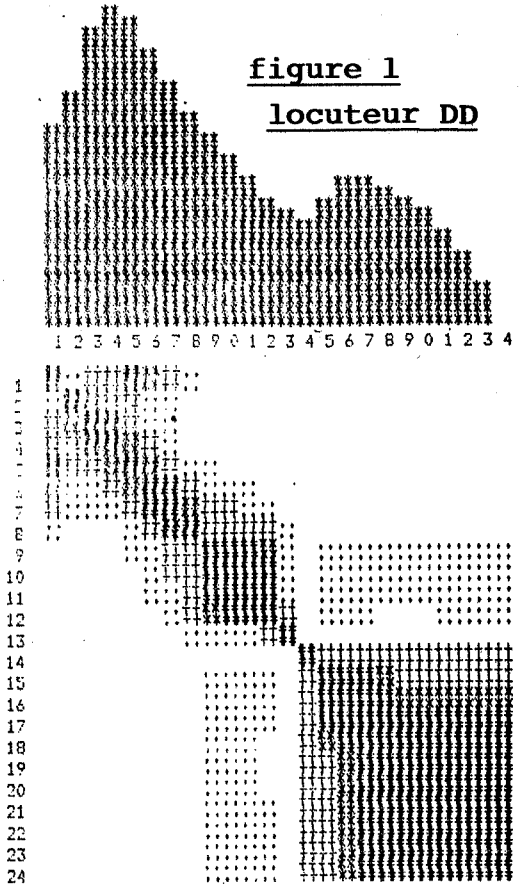
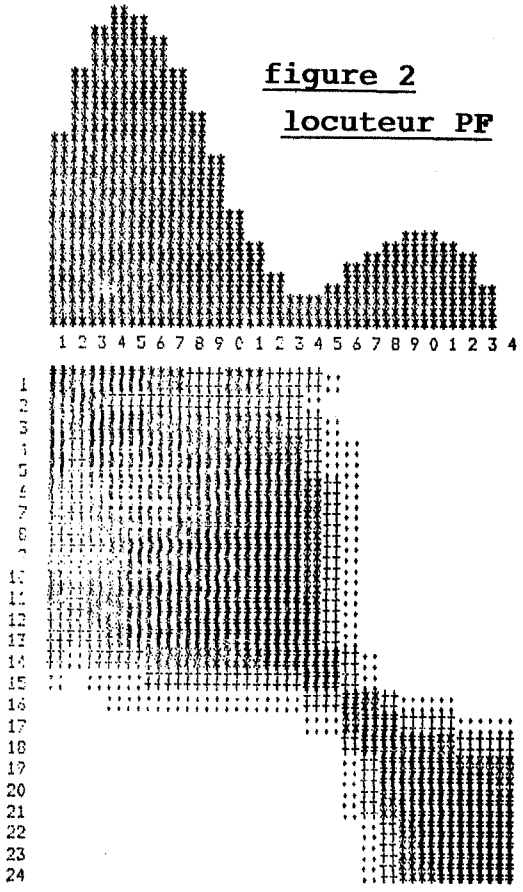


figure 2
locuteur PF

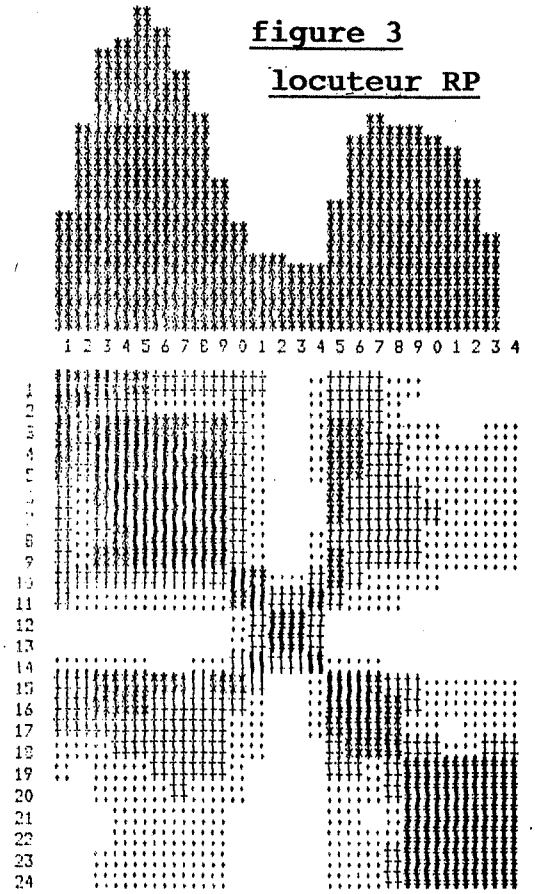


ANNEXE

Les figures 1 à 3 présentent dans leur partie supérieure, le tracé des 24 canaux du spectre et dans leur partie inférieure, la matrice de corrélations du phonème /i/.

La perspective est interlocuteur.

figure 3
locuteur RP



ANNEXE (suite)

La perspective des figures 4 et 5 est intralocuteur. La configuration des schémas est la même que pour les schémas précédents.

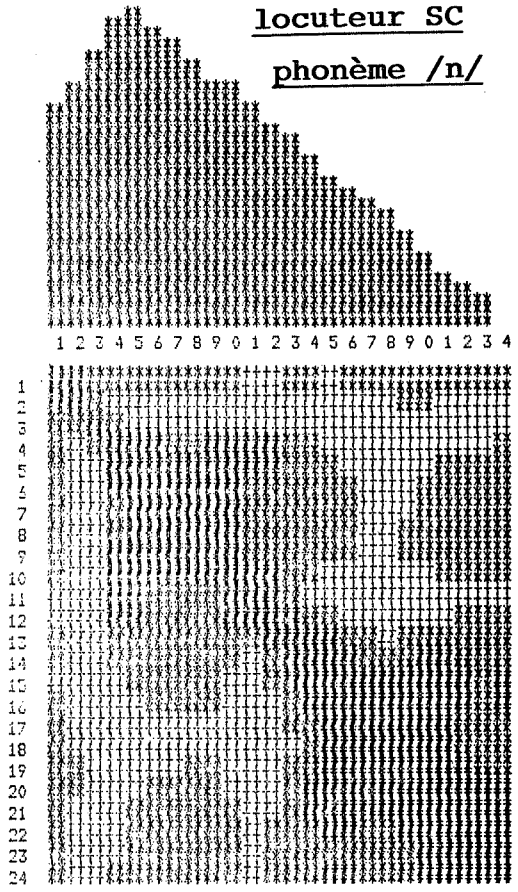


figure 4

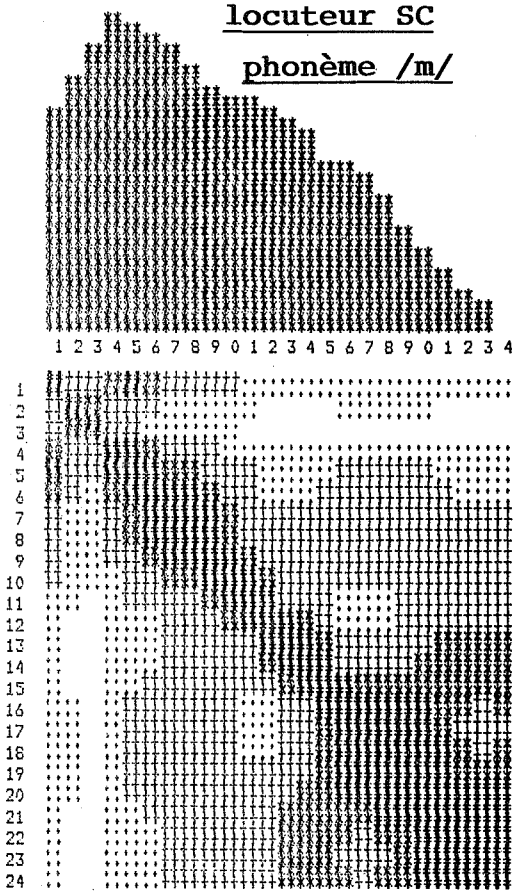


figure 5

SUR L'INVARIANCE VOCALIQUE EN FRANCAIS

M. ESKENAZI : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France.

Derrière chaque effort de caractérisation automatique de la parole, sousjacente aux expériences menées pour prouver la validité de telle ou telle approche, se trouve la vaste et très riche variabilité de la forme du message à reconnaître. La reconnaissance de la parole peut, d'une manière simple, être décrite comme la mise en relation du signal acoustique (une entité physique de nature continue) et du message (représenté par une suite d'éléments linguistiques discrets). Les premiers travaux en reconnaissance mettaient l'importance sur l'un des trois maillons, le signal (figure 1).

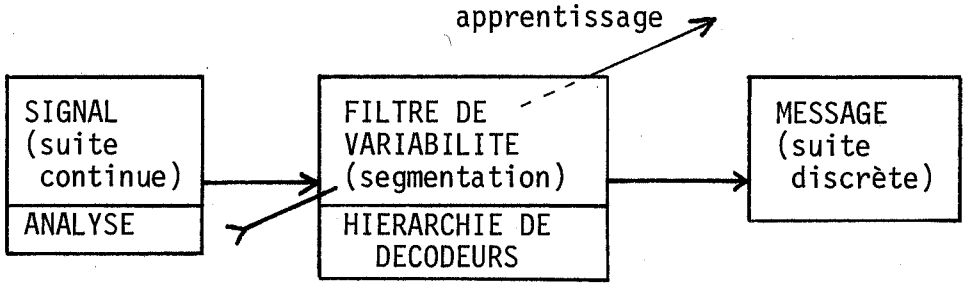


Figure 1. Schéma général de reconnaissance de la parole

L'on cherchait à retrouver un élément dans ce signal auquel correspondrait un élément du message. L'importance de la relation entre le signal et le message n'était pas mise en valeur lors de cette approche, et la nature continue du signal n'était guère prise en compte. Cette relation un à un, lorsqu'elle était appliquée à la parole dans des tâches de reconnaissance, donnait des résultats très médiocres. Ceci était le cas non seulement pour une tâche multilocuteur, mais également face à des variations intralocuteurs (1). La première réaction à l'insuffisance de cette représentation fut une recherche de l'autre côté de la chaîne de reconnaissance - une mise en valeur des éléments du message. Ici le rôle des prédictions venant des niveaux lexical, syntaxique, sémantique, et pragmatique était de contrepeser les erreurs venant du niveau acoustique (2). Une meilleure définition des éléments du message a été obtenue grâce à ces travaux, mais la troisième partie à prendre en considération, la mise en relation du signal et du message, n'a pas été adressée. Cependant, l'importance de celle-ci commençait à

être reconnue (3). Un effort soutenu se portait aussi sur l'analyse du signal. Remettant l'importance du côté signal, ces travaux portaient des fruits tels l'analyse par codage prédictif (LPC), l'analyse cepstrale, les modèles de production, et d'oreille. Le signal était alors mieux caractérisé,comme le message, de son côté, était mieux décrit, mais la complexité propre à trouver des liens entre ces deux parties n'avait pas encore été explorée.

Cette partie de mise en relation entre le signal et le message est celle qui se charge de "filtrer" la variabilité du signal de parole. Une définition de ce processus demande donc des précisions préalables sur la variabilité et donc sur le sujet de cette table ronde : l'invariance. La variabilité du signal de parole est causée par un grand nombre de faits qu'il est plus facile de dénombrer qu'il n'est de décrire leurs manifestations dans le signal. Ces "causes" fournissent des informations extra-linguistiques et sont souvent classées en quatre catégories (4) : variations intralocuteurs, variations interlocuteurs, variations dues aux conditions d'enregistrement, et variations dues à l'environnement de la parole continue. Toutes ces variations reviennent à déformer d'une manière ou une autre le message sans pour autant le rendre incompréhensible. Une déformation n'affecte pas une partie isolée du signal, mais a un effet sur le tout. La prise en compte de la variabilité (la détermination de l'invariance) doit donc tenir compte d'une relation d'ensemble et chaque "décodeur" que l'on emploie pour passer du signal au message ne saura être complètement efficace que quand il est employé en relation avec tous les autres "décodeurs" qui peuvent entrer en jeu. Pour passer du signal au message alors, et donc pour traiter l'invariance dans le signal, il faut trouver des relations entre le signal et le message, ces "décodeurs", et déterminer la relation entre ceux-ci.

Afin de définir les "décodeurs" et de faire des essais sur la hiérarchisation de ceux-ci, une quantité relativement restreinte de travaux ont été entrepris (5, 6, 7). Les "décodeurs" permettant de passer du signal au message restent liés à la nature même du signal. Ces essais ont la valeur de faire fonctionner un ensemble, un "filtre" de variabilité et peuvent donc subir des changements et servir de terrain d'essai à des décodeurs et/ou hiérarchies différents ou supplémentaires. Ces essais sont essentiels actuellement car, comme il a été remarqué lors du séminaire récent sur l'invariance (8), il n'est pas possible actuellement de dégager un consensus sur la nature de ce qui constitue ce filtre de variabilité. La seule manière de prouver la validité d'une approche automatique donnée est, en l'utilisant sur la parole, d'obtenir un bon taux de reconnaissance.

L'EVALUATION DES ELEMENTS DU FILTRE DE VARIABILITE

Qu'est-ce alors qu'un bon taux de reconnaissance ? Que cherchons-nous à

produire au niveau automatique ? A la sortie des traitements automatiques on s'attend à obtenir le message qui a été émis, un message pouvant être prononcé par n'importe quel locuteur avec toute la variation propre à chacun. Ce message peut être capté par divers types de matériel et peut être constitué de n'importe quelle suite acceptable dans la langue en question. Ceci est le but final. Mais on ne peut pas s'attendre à ce genre de résultat pour évaluer des étapes intermédiaires pour affiner des décodeurs ou pour évaluer une partie locale de la hiérarchie entre décodeurs. A ce niveau, il est nécessaire de constituer un corpus mettant en valeur le ou les décodeurs ou les relations à évaluer. Le mot "corpus" ne doit pas être compris comme étant simplement une série d'enregistrements. Il est essentiel, en établissant un corpus, que celui-ci comprenne une évaluation de son intelligibilité par des sujets humains (9). Cette évaluation, pour être utile, devrait comprendre non seulement un taux global d'intelligibilité, mais aussi les confusions faites entre différents éléments du message. Un décodeur donné peut alors être testé sur ce corpus et sa validité peut être jugée par la proximité entre ses résultats et les résultats humains, car les confusions laissées par un décodeur donné, si elles sont les mêmes, et dans les mêmes proportions que les humains, ne reflètent pas une incapacité de filtrer la variation, mais des ambiguïtés qui seront filtrées à l'aide de l'utilisation d'autres décodeurs, ou d'une autre partie de la hiérarchie.

Malgré le fait que nous ne savons pas, pour l'instant, comment l'être humain traîte la variabilité, nous pouvons souhaiter que la machine puisse obtenir les mêmes résultats que l'homme, sans prétendre que nos décodeurs automatiques soient les mêmes que ceux des humains. Il n'est cependant pas exclu que ceux-là soient inspirés de nos quelques connaissances en perception, telles l'utilisation d'une représentation logarithmique de l'amplitude et l'emploi d'une échelle de Bark ou de Mels. Il n'est pas, non plus, exclu que les résultats pour un décodeur donné puissent être supérieurs aux résultats humains (avec la réserve toutefois que les confusions restantes soient compatibles avec les confusions humaines).

INVARIANCE ET APPRENTISSAGE

Pour l'instant nous n'avons pas traité une autre variabilité qui demande une souplesse toute autre de la part de notre filtre : la variabilité de l'expérience linguistique intra-individuelle. Celle-ci doit être prise en compte comme une dimension du "filtre". Il faut rendre compte du fait qu'au fur et à mesure que nous rencontrons de nouveaux éléments linguistiques, nous les apprenons, modifiant ainsi notre filtre de variabilité. Il doit donc y avoir des mécanismes d'apprentissage. Ces mécanismes se chargent de deux actions : décider ce qui doit être appris, et alors faire les modifications correspondantes du filtre. Puisqu'une

modification, dans ce système, ne pourrait pas avoir qu'un effet local, elle entraînerait plutôt des changements compensatoires sur plusieurs décodeurs, ou sur une partie de la structure hiérarchique. L'apprentissage doit avoir pour effet l'amélioration de la reconnaissance due à l'élargissement du champ de variabilité sur lequel le "filtre" est capable d'agir.

INVARIANCE ET SEGMENTATION

Pour passer de la suite continue du signal à la suite discrète du message, il faut que les décodeurs du filtre agissent sur une suite segmentée. Il est peu probable que les segments pertinents pour les différents décodeurs soient tous de la même taille physique. Si par exemple (et comme nous le prouverons plus loin), des segments stables du signal peuvent être caractérisés en agissant sur un segment d'assez courte durée, les parties transitoires demandent un suivi de leur évolution temporelle et un segment bien plus long est nécessaire. Les dernières années ont vu un foisonnement de travaux utilisant des segments de base tels que la demi-syllabe, la syllabe et le mot. Mais jusqu'à ce jour il n'y a pas à notre connaissance, de travaux qui emploient plusieurs tailles de segments selon la nature du segment et du décodeur.

INVARIANCE ET ANALYSE

Jusqu'ici nous avons essayé de faire une esquisse de ce qui devrait théoriquement se passer entre le signal et le message. Nous allons maintenant donner un exemple concret d'un décodeur possible (10), son évaluation, et la preuve de sa robustesse vis-à-vis de l'analyse (prouvant que le filtrage de la variation ne comprend pas l'analyse, celui-ci faisant partie plutôt du premier maillon de la chaîne de reconnaissance : le signal). Nous décrivons d'abord notre démarche, ensuite les tests, et finalement les remarques sur la variabilité qui ressortent à partir de cette évaluation.

L'extraction de la courbure

Nous avons développé un décodeur agissant sur les parties relativement stables du signal qui permet, pour l'instant, la reconnaissance de douze voyelles orales et nasales du français. Le corpus utilisé pour l'élaboration du décodeur comprenait 50 ms. de la partie stable des voyelles se trouvant dans une phrase cadre "J'ai dit t - six fois", prononcée par trente locuteurs masculins et féminins. Les tests qui sont décrits plus loin se basent sur un deuxième corpus, composé de trente locuteurs masculins et féminins dont dix ont été utilisés pour constituer la base statistique, et vingt pour les tests de reconnaissance. C'étaient des voyelles prononcées isolément dont nous avons extrait nos segments de 50 ms.

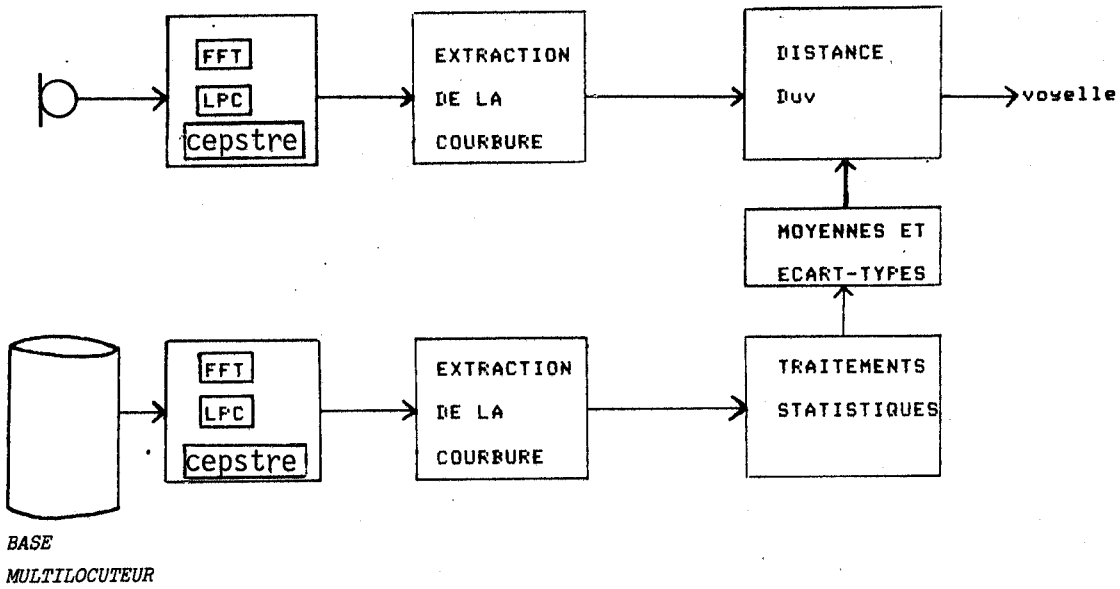


Figure 2 - Synoptique de la démarche

La figure 2 donne le synoptique de notre démarche. Pour créer une base statistique, les voyelles des dix locuteurs ont été traitées de la manière suivante. Chaque segment de 50 ms, après une fenêtre de Hamming et une préaccentuation, a été analysé, soit par une FFT, soit par codage prédictif, soit par analyse cepstrale (la même analyse a été employée pour tout le corpus, ceci donnant donc lieu à trois tests différents de reconnaissance, un pour chaque type d'analyse. Une fois que le signal a été analysé, les démarches propres au décodeur ont été entreprises. D'abord, nous avons regroupé les canaux, linéairement répartis en fréquence après analyse, en 32 canaux séparés selon une échelle de Bark. Ensuite, la valeur de l'amplitude a été transformée selon une échelle logarithmique (sauf dans le cas de l'analyse cepstrale où cette transformation est effectuée lors de l'analyse). La prochaine étape est celle que nous appelons "extraction de la courbure" (et qui ne correspond pas à la conception proprement mathématique de la courbure). Cette étape consiste en deux parties : d'abord une série de lissages sévères sur le spectre, ce qui enlève des informations non-pertinentes et donne une certaine indépendance par rapport à la variabilité locale intralocuteur ; ensuite un spectre lissé est soustrait d'un autre, moins lissé, afin d'obtenir une indépendance de l'amplitude et donc du "spectral tilt", ou la distribution relative de l'énergie le long du spectre. Ceci donne une prise en compte de la variabilité interlocuteur.

La prochaine étape, celle de la représentation statistique de cette base multilocuteur, présume une distribution gaussienne des données. Pour chaque voyelle nous n'avons gardé alors que deux vecteurs de 32 valeurs : le premier vecteur

représente la valeur moyenne de la "courbure" de la voyelle à chaque canal ; et le deuxième vecteur, les valeurs de l'écart-type.

La première ligne de la figure 2 représente la démarche des tests de reconnaissance. Chaque segment de 50 ms provenant des voyelles des 20 locuteurs de test a subi les mêmes analyses que les segments utilisés pour la base. Un segment inconnu a alors été comparé aux références de base par une distance, D_{uv} :

$$D_{uv} = \frac{|C_u - m_v|}{\sigma_m}$$

où C_u est la valeur de courbure de la voyelle inconnue au canal en question, m_v et σ_m la moyenne et l'écart-type de la voyelle de référence à chaque canal, K, du spectre. Les distances pour tous les canaux par rapport à une voyelle donnée ont ensuite été sommées :

$$\mathcal{D}_{uv} = \sum_{K=1}^{32} D_{uv}(K)$$

et la voyelle ayant la distance minimum avec le segment inconnu a été proposée comme celle qui a été reconnue.

Evaluation du corpus

Afin de valider le corpus multilocuteur de voyelles isolées et d'établir une base d'évaluation des résultats automatiques, nous avons procédé à des tests d'intelligibilité du corpus. Treize sujets ont entendu ces voyelles, notant leurs réponses dans une forme graphémique proche de l'Alphabet Phonétique International à laquelle ils avaient été entraînés. Le résultat de leurs réponses, comportant les confusions constatées, est visualisée sous forme de la matrice de confusion de la figure 3. Les voyelles en début de rang représentent celles que les locuteurs entendaient prononcer et celles en tête de colonne, celles que les sujets ont reconnues. La diagonale fortement suivie souligne l'intelligibilité du corpus. Cependant quelques départs de cette diagonale sont à remarquer : des paires de voyelles en évolution phonologique actuelle (oe/ø, ɛ/e) et des voyelles nasales. Seulement les taux de confusion de 10% ou plus ont été rapportés ici pour plus de clarté. Le taux global d'intelligibilité est de 77,9%, ce qui est comparable à d'autres tests (11, 12) pour une tâche de difficulté comparable.

Cette matrice peut alors servir de moyen d'évaluation des résultats des tests de reconnaissance automatique si nous en soustrayons la matrice des résultats d'un test de reconnaissance donné.

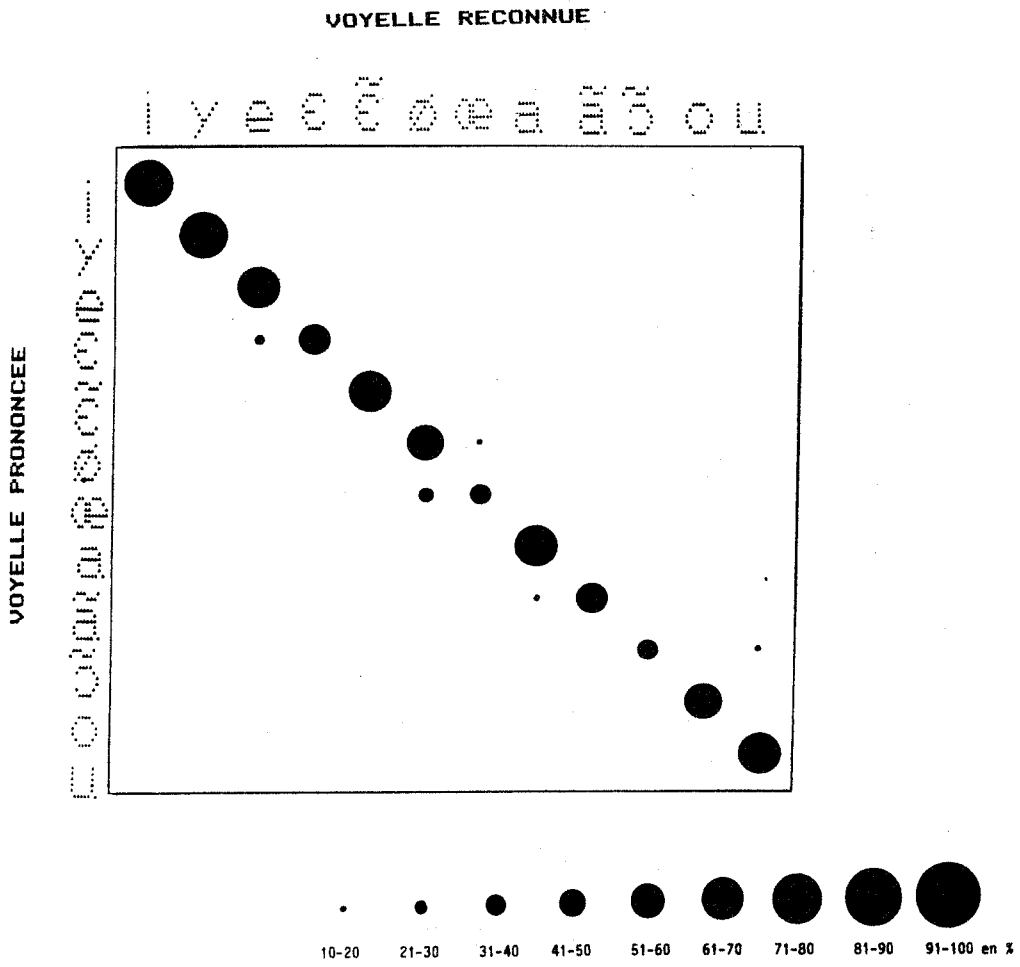


Figure 3 - Résultats du test d'intelligibilité
exprimé par une matrice de confusion

Résultats et leur relation à notre définition du "filtre"

Les figures 4 et 5 montrent les matrices de confusion pour une analyse FFT et une analyse par codage prédictif respectivement, les résultats complets pour l'analyse cepstrale n'étant pas disponibles au moment de l'écriture de cet article. A première vue, il est évident que l'analyse par codage prédictif donne de meilleurs résultats (76,3% globalement, donc moins de 2% de différence avec les résultats humains). Mais ceci n'est pas surprenant au vu des confusions issues d'une analyse par FFT car la grande majorité d'entre elles (voir le cas du /u/, par exemple) proviennent du fait que le fondamental de la voix est encore présent dans le spectre et recouvre ou cache d'autres informations qui sont importantes pour la reconnaissance des voyelles. En ce qui concerne l'analyse par codage prédictif, par contre, l'on s'attend à un grand nombre de confusions centrées sur les voyelles nasales dûes aux problèmes bien connus de la prise en compte des zéros dans le spectre de la LPC. Les confusions sont cependant les mêmes que pour l'analyse par FFT. Ceci est dû aux lissages sévères opérés sur le spectre lors de

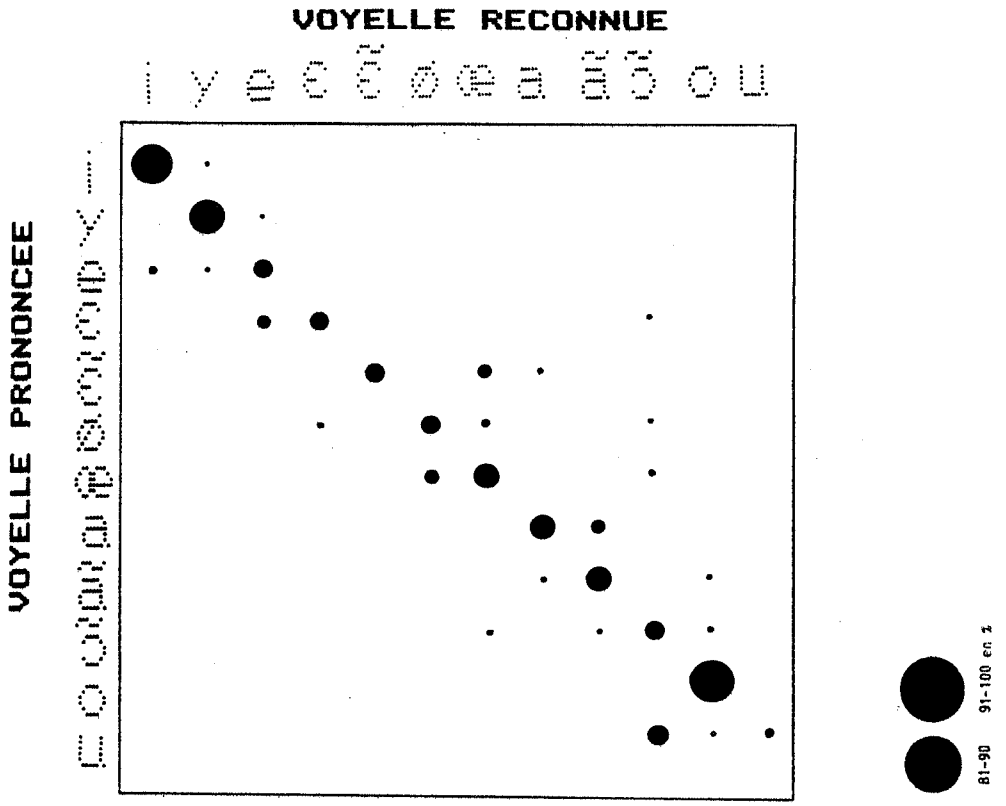


Figure 4 - Résultats obtenus avec une analyse spectrale par FFT

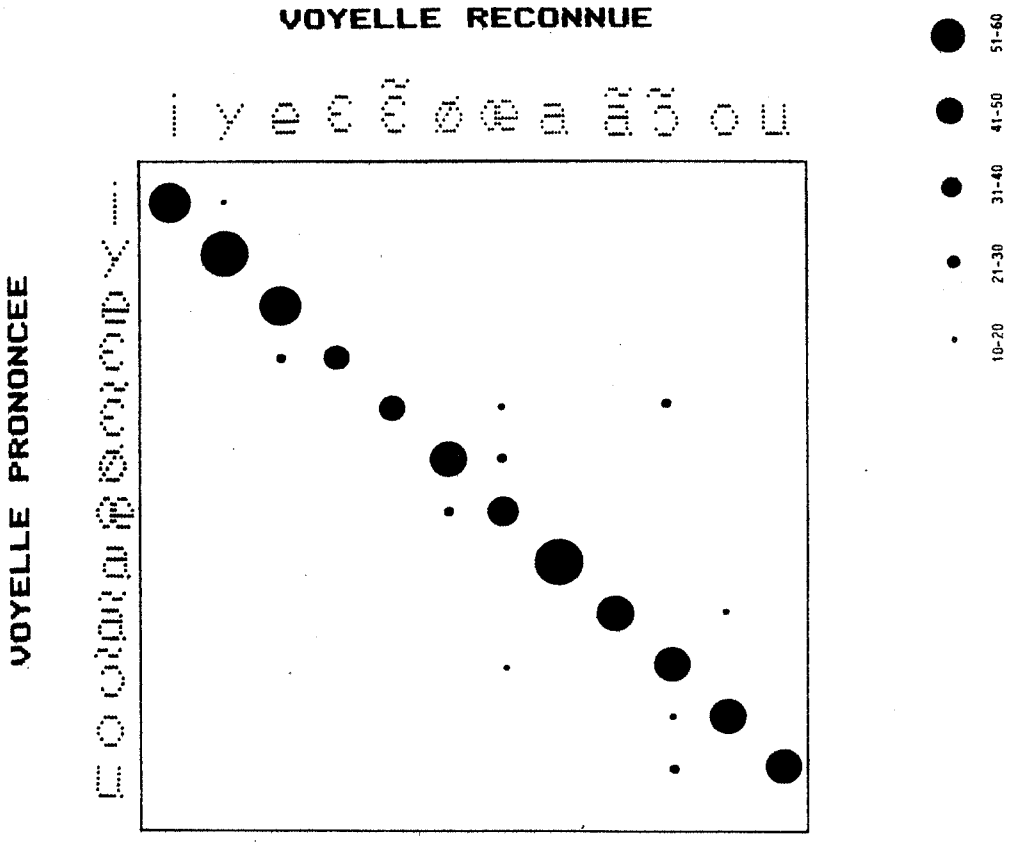


Figure 5 - Résultats obtenus avec une analyse spectrale par codage prédictif

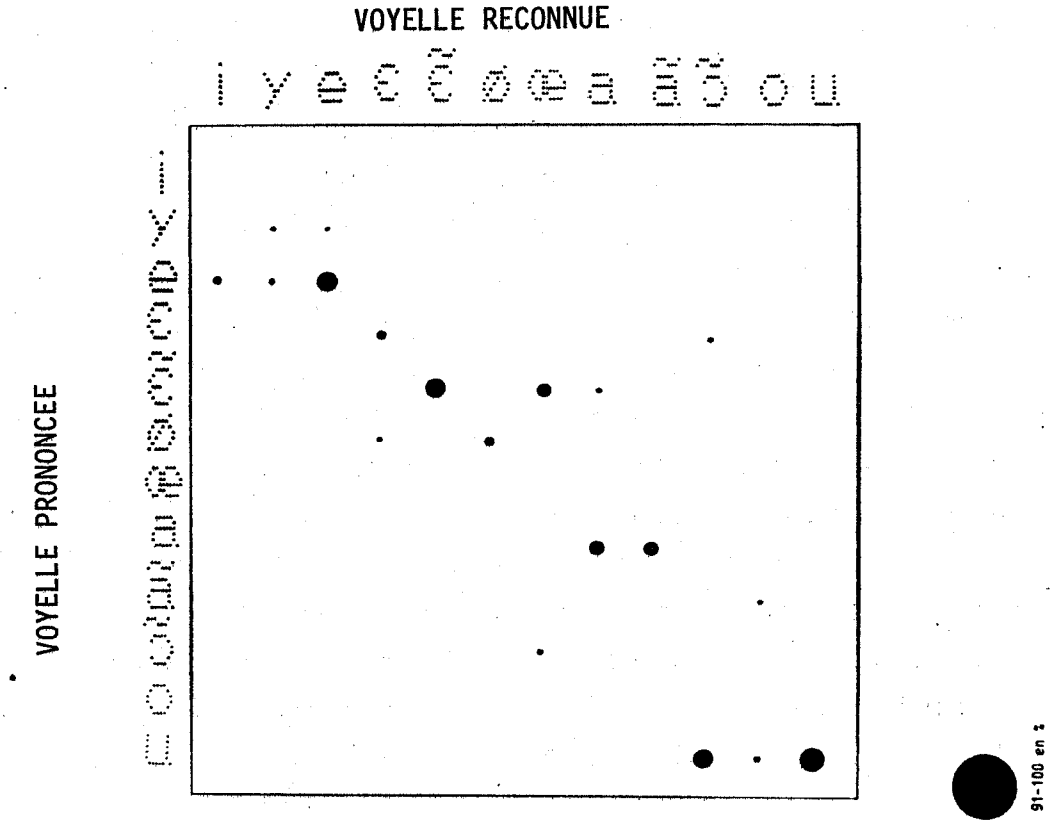


Figure 6 - Différences résultats FFT/résultats humains

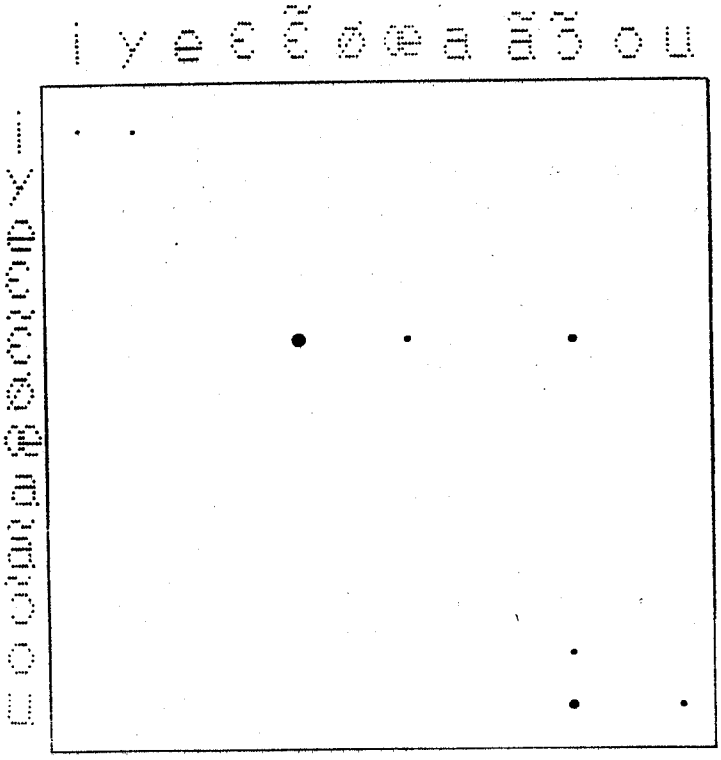


Figure 7 - Différences résultats LPC/résultats humains

l'extraction de la courbure. Les figures 6 et 7 montrent la différence entre les résultats automatiques et les résultats humains pour l'analyse par FFT et l'analyse par codage prédictif respectivement. Ce qui ressort le plus de la comparaison de ces deux figures est que les mêmes confusions sont faites lors des deux analyses. Le fait d'obtenir le même genre de résultat (ce qui,nous l'espérons,sera le cas pour l'analyse cepstrale également) souligne la robustesse du décodeur et son indépendance par rapport à l'analyse spectrale. Chaque analyse spectrale offre une interprétation différente du signal ou de la production du signal et reste liée à la représentation du signal et non pas au filtre de la variabilité. Il restera à démontrer ailleurs que la différence du taux de reconnaissance entre les diverses méthodes d'analyse peut être compensée par l'action d'autres décodeurs en relation avec celui-ci (comme nous l'avons indiqué plus haut, nous nous attendons à ce que les confusions qui restent ici soient levées dans l'interaction avec le reste du "filtre"). Les confusions qui restent pour les voyelles nasales, par exemple, sont, au moins partiellement, dues à la taille du segment pris en compte. Il est probable qu'ici la demi-syllabe, ou la syllabe soient d'une taille plus adaptée à la reconnaissance de voyelles nasales.

CONCLUSIONS

Nous avons, à l'aide de l'exemple d'un décodeur de caractérisation fréquentielle grossière, essayé d'explorer quelques aspects du traitement automatique de la variabilité. Le problème de l'évaluation d'une partie de l'ensemble que nous appelons "filtre" trouve une première réponse dans l'outil fourni par une matrice de confusion représentant l'intelligibilité d'un corpus donné. Les tests ici laissent entrevoir des exemples au niveau de la segmentation du signal continu. En ce qui concerne la relation invariance-analyse, nos tests montrent que les éléments du filtre doivent être complètement indépendants des diverses méthodes d'analyse qui peuvent être employées pour décrire le signal.

La tâche de la prise en compte de la variabilité du signal est extrêmement complexe. Du point de vue des décodeurs et de leur relation hiérarchique, beaucoup de questions restent à explorer. L'apprentissage implique des structures souples et des décodeurs étroitement liés. Ce ne sera qu'une fois que des réponses à tous ces problèmes auront été trouvées que nous pourrons changer le nom, filtre de variabilité, en relation d'invariance.

REFERENCES

1. PARKER, F., "Distinctive features and acoustics cues", J. of the Acoustical Soc. of America, vol. 62, n° 4, pp. 1051-1054, 1977.

2. REDDY, D.R., "Speech recognition by machine : a review", Proceedings of the IEEE, 76CH1067-8 ASSP, Philadelphia, Pa, pp. 501-531, 1976.
3. GOODMAN, G., REDDY, R., "Alternative control structures for speech understanding systems", in Trends in Speech Recognition, ed. W. Lea, pp. 234-246, 1980.
4. KLATT, D., "The problem of variability in speech recognition and in models of speech perception", Symposium Variance-Invariance in Speech, MIT, Boston, 1983.
5. ROSSI, M., NISHINUMA, Y., TREVARIAN, O., MERCIER, G., "Reconnaissance des voyelles par les indices et les traits", Séminaire GALF-AFCET Processus d'Encodage et de Décodage Phonétiques, Toulouse, pp. 2-28, 1981.
6. ABRY, C., BOE, L.J., "Unités et niveaux de traitement du signal de parole : apports d'une analyse linguistique intrastratique", Séminaire GALF-AFCET, Processus d'Encodage et de Décodage Phonétiques, Toulouse, pp. 29-40, 1981.
7. CAELEN, J., CAELEN, G., "Indices et propriétés dans le projet ARIAL II", Séminaire GALF-AFCET, Processus d'Encodage et de Décodage Phonétiques, Toulouse, pp. 128-143, 1981.
8. Symposium Variance-Invariance in Speech - Conclusions, MIT, Boston, octobre 1983.
9. ESKENAZI, M., LIENARD, J.S., "Recognition of steady-state French sounds pronounced by several speakers : comparison of human performance and an automatic recognition algorithm", Speech Communication, vol. 2, pp. 173-177, 1983.
10. ESKENAZI, M., "Caractérisation automatique des voyelles du français en vue de leur reconnaissance automatique", Thèse de Toisième Cycle, Informatique, Université Paris XI, 1984.
11. GOTTFRIED, T.L., STRANGE, W., "Identification of coarticulated vowels", J. of the Acoust. Soc. of Amer., vol. 68, n° 6, pp. 1626-1635, 1980.
12. DODDINGTON, G.R., Communication personnelle, 1983.

UNE APPROCHE GLOBALISTE DE LA VARIABILITE ACOUSTICO-PHONETIQUE DE LA PAROLE

J.S. LIENARD : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France.

I. INTRODUCTION

La variabilité du signal de parole, est le problème central de l'analyse, qu'elle soit vue sous l'angle de la perception humaine ou sous celui de la reconnaissance automatique. La variabilité se manifeste dans de nombreux aspects du message. Elle est particulièrement évidente sous l'aspect phonétique. Lorsque sont apparus, il y a une quarantaine d'années, des modèles - mathématiques puis électriques - du conduit vocal, et des moyens d'investigation raffinés comme l'oscilloscope et le spectrographe, on a cru pouvoir caractériser les phonèmes par des propriétés acoustiques absolues déduites de l'analyse de segments contigus - spectres, formants, loci des formants. Lorsqu'il devient évident qu'une telle correspondance entre sons et symboles phonétiques était illusoire, on cherche des grandeurs intermédiaires, des traits distinctifs, qui semblaient, du point de vue phonologique, constituer une base de décomposition universelle de l'ensemble des phonèmes (1). Mais, ici encore, il s'est avéré impossible d'associer de manière certaine des mesures acoustiques à chaque trait distinctif. La simplification apportée par le système de traits distinctifs n'est effective que sur le plan phonologique, mais n'a pas fait évoluer le problème de la variabilité acoustique : il est aussi difficile d'associer à un segment de signal les traits "vocalique", "compact", ... que d'y associer le symbole /a/, et ce pour des raisons semblables : variation des mesures acoustiques selon le contexte phonétique, le type de voix, le locuteur.

Il est vraisemblable qu'une démarche encore plus analytique - par exemple décomposition de chaque trait en "indices acoustiques", de chaque indice en "propriétés" (cf ABRY et BOE, dans (2)), etc - se heurtera à un obstacle du même genre. Le choix des primitives est arbitraire dans chaque cas ; rien ne garantit leur indépendance ou leur pertinence phonétique. Cependant une telle approche, si elle est suivie en tenant compte des expériences psycho-acoustiques, peut peut-être aboutir à un terrain plus ferme, c'est-à-dire à des éléments incontestables du point de vue de la perception. Par ailleurs ce type de démarche a le mérite de montrer que ce n'est pas dans des mesures acoustiques plus ou moins complexes qu'il faut chercher des invariants, mais dans la manière dont les mesures acoustiques élémentaires sont structurées, avec sans doute plusieurs

niveaux de structuration entre le signal et la séquence de symboles linguistiques qui lui est associée.

Si l'on considère des entités linguistiques plus grandes que le phonème - syllabe, mot - on rencontre encore le même problème ; cette fois la variabilité porte sur plusieurs segments successifs et aussi sur l'échelle de temps. Il est d'autant plus difficile de comparer plusieurs séquences de même contenu linguistique mais provenant de plusieurs locuteurs différents. Les méthodes usuelles de reconnaissance par mots, qui ne font que compenser les différences d'échelle temporelle, sont en échec dès que le locuteur change légèrement sa voix, ou dès qu'on change de locuteur. Comme il s'agit de phénomènes complexes, comportant de nombreuses dimensions avec, chacune, un degré de liberté, un simple moyennage de plusieurs occurrences n'améliore pas la qualité des entités de référence, bien au contraire. Une manière de contourner ce problème, en reconnaissance par mots, est d'utiliser plusieurs références obtenues par agglutination ("clustering"), et donc représentant au mieux un large échantillon d'apprentissage. Cependant cette technique ne nous apprend rien sur la structure acoustique des mots considérés, et possède les mêmes limitations que la reconnaissance par mots, avec des algorithmes plus gourmands en temps et en mémoire.

Notre thèse est qu'aux entités phonétiques et phonologiques identifiées par les linguistes correspondent bien des invariants acoustiques, mais que ceux-ci sont à rechercher dans les relations structurales qu'entretiennent entre eux les paramètres acoustiques, et non dans les valeurs des paramètres eux-mêmes. La variabilité est le problème dual de l'invariance. S'il existe des invariants acoustiques structuraux à signification phonétique, alors, dans une large mesure, les variations observées peuvent elles aussi constituer des systèmes de signes et transmettre des informations d'une autre nature (prosodique, diagnostique).

Exemple : ceci est un bâton 

ceci est aussi un bâton 

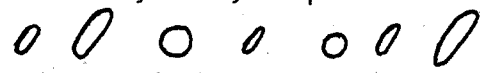
ceci est un rond 

ceci est un autre rond 

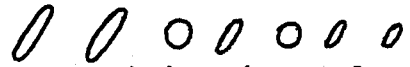
Je peux faire un message de bâtons et de ronds signifiant "Bonjour" en Ronbaton 

Voici une autre version du même message, provenant de mon ami Paul :

Ma cousine Berthe, elle, le prononce comme ceci quand elle est gaie :



ou comme cela quand elle est triste :



Dans la mesure où je sais que les invariants linguistiques du Ronbaton sont des ronds et des bâtons, je peux identifier les éléments transmis, même si les ronds de Berthe ressemblent aux bâtons de Paul et réciproquement. Je peux aussi m'intéresser à leurs variations, et, sans connaître le message transmis, supposer que

vient de Berthe ; et je peux même inférer qu'elle est de bonne humeur, et que le "bonjour" de Paul trahit des pensées lugubres ... Bien sûr, je peux me tromper, et si Paul m'affirme qu'il exprime ainsi une joie sans réserve, je saurai que

transmet en fait les mêmes informations que le message précédent de Berthe.

II. SUR LA VARIABILITE

S'il est couramment admis que la variabilité du signal - à contenu phonétique identique - présente un aspect intra-locuteur et un aspect inter-locuteur, un troisième aspect est généralement ignoré, ou amalgamé à l'un des précédents. Il s'agit de la variabilité observée en fonction des conditions d'enregistrement (prise de son, transmission, bruit, réverbération, etc). Mais nous nous intéressons surtout, ici, à la variabilité liée à l'émission du signal, ou variabilité de source.

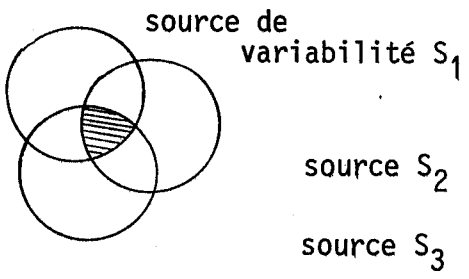
Notons tout d'abord que le terme de variabilité intra-locuteur peut être pris dans une acceptation large ou restreinte. L'acception large se formule ainsi : soit un locuteur l, prononçant une séquence de contenu linguistique donné ; la variabilité intra-locuteur large de ce locuteur VAL(l), ou son champ de liberté acoustique, est l'ensemble de toutes les manières possibles, pour ce locuteur, de prononcer cette séquence. Par exemple je pourrai prononcer "Salut les copains" avec une voix forte, moyenne, faible ou chuchotée, avec ma voix normale ou ma voix criée, chantée, etc, tout en y mettant le même contenu linguistique. Au sens restreint la variabilité sera définie de la manière suivante : soit un locuteur l qui s'astreint à prononcer une même séquence de manière reproductible, avec le même type de voix, la même intention, la même prosodie etc ; la variabilité restreinte intra-locuteur de ce locuteur VAR(l) est l'ensemble des variations produites dans ces conditions. C'est la partie du processus de génération du signal qui échappe au contrôle du locuteur. En général, c'est cette variabilité restreinte qui est présumée dans les expériences de reconnaissance

automatique de la parole. Selon la capacité du locuteur à contrôler son émission, les résultats seront plus ou moins bons. Nous connaissons tous des locuteurs qui ont de bons résultats même avec des systèmes de reconnaissance médiocres, et réciproquement.

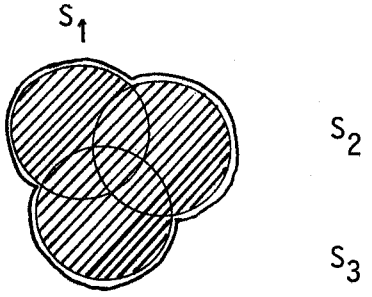
La définition de la variabilité inter-locuteur est liée à la précédente. Si l'on demande à tous les locuteurs considérés, constituant une population p, d'imiter au mieux un même modèle on se place sur le plan de la variabilité inter-locuteur restreinte VRR(p) de cette population. Il est bien rare que tous les locuteurs de la population puissent fournir une réalisation identique. Donc en général VRR(p) est constitué d'îlots disjoints. Si l'on s'attache à recueillir un échantillonnage aussi large que possible de leurs productions, à contenu linguistique identique, on se place sur le plan de la variabilité inter-locuteur large VRL(p) de cette population.

Dans les expériences de reconnaissance multi-locuteur, on adopte en général un point de vue intermédiaire : la consigne implicite faite à chaque locuteur est d'adopter une voix "standard", moyenne, reproductible donc dans l'optique VAR(1); et l'on considère en extension la population p, donc dans l'optique VRL(p).

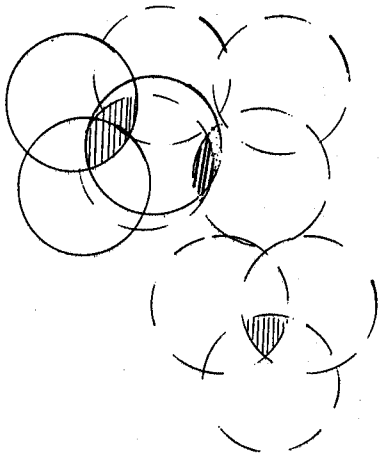
Ces propos peuvent être illustrés de la manière suivante :



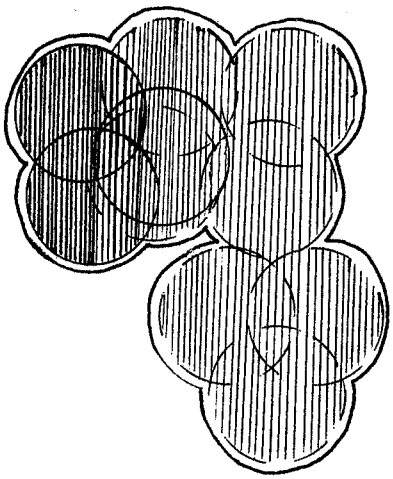
Variabilité intra-locuteur restreinte, pour un locuteur : c'est la variabilité résiduelle lorsque le locuteur s'efforce de parler de manière reproductible, c'est-à-dire de compenser toutes les sources de variabilité S : (timbre, force de voix, rapidité d'élocution, etc), à contenu linguistique donné



Variabilité intra-locuteur large, pour un locuteur : c'est l'union de toutes les sources de variabilité, à contenu linguistique donné



Variabilité inter-locuteur restreinte, pour une population de trois locuteurs : c'est l'union des variabilités restreintes des divers locuteurs (zones hachurées)



VRL(p)

Variabilité inter-locuteur large : c'est l'union des variabilités larges des divers locuteurs

La notion de variabilité, telle que nous l'avons évoquée jusqu'à présent, a quelquechose d'insatisfaisant. C'est qu'elle suppose que l'on ne s'intéresse qu'à un seul aspect du message à la fois. Alors les variations, qui sont peut être porteuses d'un autre type d'information comme nous l'avons vu dans la section I, sont considérées comme des parasites, qu'il faut chercher à éliminer ou à compenser. Mais est-il possible, et raisonnable, de supposer que le système récepteur - homme ou machine - ne puisse saisir qu'un seul type d'information à la fois ? Si l'on admet le contraire, alors il faut abandonner la notion de variabilité, du moins au sens large. En effet, cela signifie que la variabilité observée sous un aspect peut en fait être le support d'une information pertinente sous un autre aspect. Les deux aspects peuvent être en interaction, si bien que l'on ne peut décoder l'un sans l'autre.

Prenons quelques exemples concrets. La prosodie transporte une information suprasegmentale, dont la nature n'est pas encore exhaustivement explorée. La prosodie se manifeste acoustiquement à travers plusieurs paramètres : répartition des pauses, évolution de la hauteur, de la durée, de l'intensité. Si la hauteur peut être évaluée en première approximation par des mesures acoustiques relativement simples (fréquence laryngienne), il semble évident que l'évaluation des

durées phonémiques ou syllabiques requiert la connaissance préalable des segments considérés - phonèmes ou syllabes - et même leur identification. Réciproquement la connaissance de la durée d'un phonème est quelquefois nécessaire pour l'identifier. Voici donc deux aspects - décodage phonétique et décodage prosodique - en interaction étroite et évidente.

Prenons un autre exemple. Le VOT des consonnes occlusives sourdes a fait couler beaucoup d'encre. S'il est bien établi que, dans des conditions de variabilité intra-locuteur restreinte, le VOT permet à lui seul de séparer trois catégories de consonnes, il n'en va plus de même si l'on se place dans d'autres conditions de variabilité. En particulier cet indice est très sensible au type de voix adopté par le locuteur ainsi qu'à la vitesse d'élocution, l'intensité, etc. Pour que l'information du VOT puisse être pleinement exploitée dans un processus de reconnaissance, il faudrait donc être capable d'identifier le locuteur, son type de voix, sa rapidité d'élocution, etc.

Une telle situation se retrouve pratiquement pour chacun des indices acoustiques dégagés par l'analyse. Soit, par exemple, un maximum observé à un instant donné dans le spectre. Sans même chercher à donner à ce maximum la dénomination de Formant, il est essentiel, pour pouvoir exploiter son information, de savoir s'il est lié au signal ou non (il peut, par exemple, être dû à la fonction de transfert du canal de transmission, ou à un bruit parasite) et, dans le premier cas, s'il est lié au fondamental de la voix, ou au timbre particulier du locuteur, ou, ce qui est implicitement posé, aux phonèmes émis. Et même dans ce dernier cas, on ne peut être certain de la pertinence phonétique de cet indice acoustique, qui n'a de valeur que comme partie d'un ensemble d'indices avoisinants.

Nous pensons donc

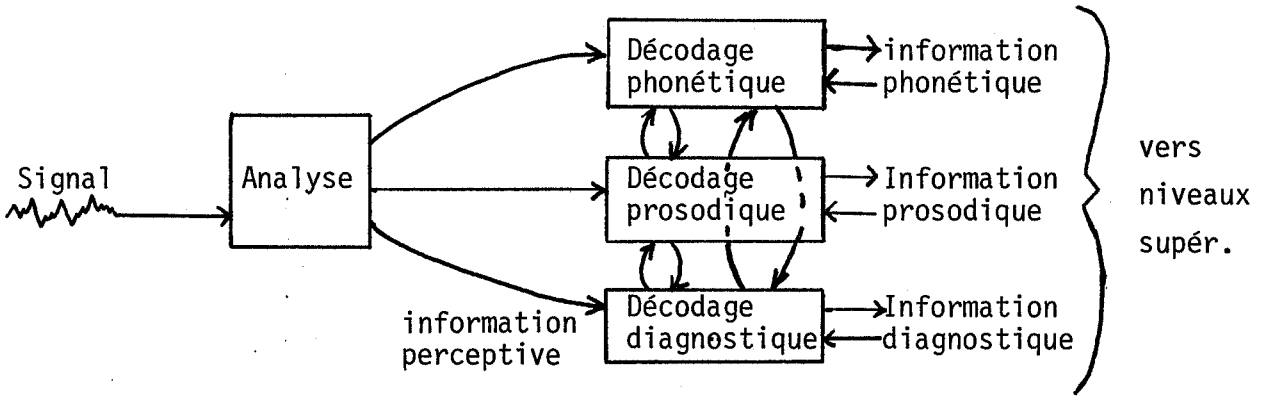
1) que la variabilité, du moins au sens large, n'est qu'une mauvaise excuse à notre ignorance des structures du signal. Dans le domaine acoustico-phonétique, les présentations du type "nous avons découvert de bons indices, malheureusement il y a quelques exceptions" abondent

2) que les difficultés rencontrées dans le décodage automatique viennent du fait que l'on suppose que les réalisations des diverses informations convoyées par le signal sont indépendantes. Notre schéma mental du décodage acoustico-phonétique ressemble à celui d'un système linéaire : à une entrée E_1 acoustique correspondrait une sortie S_1 phonétique ; à E_2 correspondrait S_2 diagnostique, à E_3 correspondrait S_3 phonétique, etc. Et à la somme ΣE_i correspondrait ΣS_i. Ceci est sans doute faux.

Pour prendre en compte les remarques précédentes, il nous faut considérer que tous les aspects du signal doivent être soumis simultanément à l'analyse.

Au lieu de supposer chaque paramètre indépendant, et s'apercevoir ensuite qu'il n'en est rien dès lors que l'on considère des signaux réels, il semble préférable de considérer a priori que tous les aspects sont liés, quitte à découvrir que parfois ils sont indépendants. Ceci implique un changement dans les méthodes d'analyse de la parole, que nous allons aborder dans la section III.

Dans le domaine du décodage acoustico-phonétique, une approche prenant en compte les diverses informations présentes dans le signal peut être schématisée comme suit



Ce schéma signifie que, a priori, chaque maillon de décodage tient compte des autres maillons et des informations fournies par les niveaux supérieurs. Il est simplifié, dans la mesure où le maillon d'analyse est considéré comme indépendant des informations efférentes, et dans la mesure où les maillons de décodage sont représentés comme des traitements opérant à un seul niveau, alors qu'ils en comportent vraisemblablement plusieurs.

III. A PROPOS DE L'ORGANISATION DES DONNEES SENSIBLES

Actuellement il existe deux méthodes pour affecter une étiquette à un ensemble de données provenant du monde extérieur. La première est une comparaison globale de cet ensemble à des ensembles de données de référence, mémorisées au préalable et rendues plus ou moins significatives par d'éventuels traitements statistiques. C'est typiquement l'approche suivie en reconnaissance par mots, et les limites en sont évoquées plus haut. La seconde approche consiste à décomposer les données en éléments, et à définir chaque séquence comme une combinaison de ces éléments par exemple au moyen d'un ensemble de règles de réécriture. L'opération de reconnaissance implique alors la détection et l'identification des éléments, et l'examen de leur organisation comparée à celles qui ont été mémorisées par le biais des règles.

En dépit de sa puissance apparente et de ses nombreuses variantes selon les

grammaires utilisées, la seconde méthode (reconnaissance syntaxique) se heurte à de nombreuses difficultés. La détermination des éléments - ou primitives - est souvent arbitraire ; l'ensemble des règles, établies manuellement par un "expert" au vu d'un échantillon des données - peut rarement excéder un certain degré de complexité car il est difficile d'en prévoir toutes les combinaisons et d'éviter les bouclages et les contradictions. Et surtout, on ne sait pas expliciter, en matière de parole, les aspects du signal dans lesquels se trouve l'information que l'on cherche à définir. En l'état actuel des recherches, les approches de ce type ne peuvent fournir plus que la connaissance explicitement introduite dans le système, ce qui entraîne de très fortes contraintes sur les données traitées.

D'une certaine manière, nous sommes dans la même situation, avec nos méthodes automatiques de reconnaissance, que les psychologues du début du 20^{e} siècle avec les doctrines relatives à la perception humaine, essentiellement axées sur la recherche de sensations élémentaires et sur leur association. Bien entendu nos moyens de description des associations sont infiniment plus puissants que celles auxquelles on pensait à l'époque, et qui se limitaient à une simple proximité temporelle ou spatiale des sensations élémentaires. Il n'en reste pas moins que nous cherchons à définir des constituants élémentaires et des structures, sur des modèles inspirés de la physique des atomes et des molécules. Mais il est bien difficile de discerner dans le signal de parole des éléments stables et incontestables ; difficile de dégager des structures bien nettes, du moins aux niveaux infra-phonémiques, et sous les aspects prosodiques et diagnostiques ; difficile également de définir des lois d'auto-organisation ou d'apprentissage de ces structures.

La Théorie de la Forme (Gestalt-Theorie) s'est élaborée en réaction à la psychologie analytique et associationniste du 19^{e} siècle. Nous n'en ferons pas ici un exposé complet, que l'on pourra trouver dans (3, 4, 5), mais nous en rappellerons quelques-uns des aspects essentiels.

Tout d'abord la Théorie de la Forme pose la prééminence de l'organisation d'ensemble des données sensibles, par rapport aux parties : une partie dans un tout est autre chose que cette même partie dans un autre tout.

La notion même d'élément disparaît pratiquement, selon son degré de cohésion interne, une Forme peut être forte (c'est-à-dire indissociable en parties) ou faible, c'est-à-dire décomposable en parties qui sont elles-mêmes des Formes, et qui sont articulées entre elles.

Toute Forme a une tendance naturelle à s'identifier à la Forme voisine la plus symétrique, la plus simple, la plus régulière. C'est la loi de la "bonne forme" qui, malheureusement, n'a de sens qu'à partir du moment où un espace et une métrique peuvent être définis.

Une Forme peut émerger à partir d'un groupe de sous-Formes en fonction de leur proximité et de leur ressemblance (même remarque que ci-dessus). D'une manière générale une sous-Forme s'intégrera d'autant plus facilement à une Forme existante qu'elle s'inscrit dans une continuité et ne perturbe pas les rapports existant dans la Forme initiale.

Enfin toute Forme s'inscrit sur un fond, et entretient avec lui des relations particulières, bien mises en évidence par les expériences de masquage et de perception de figures ambiguës.

La théorie de la Forme prend en compte d'emblée les informations afférentes (associationnisme, ou structuration ascendante) et efférentes (présélection par des "attitudes", au sens psychologique, ou prévisibilité, ou structuration descendante). Bien que le terme de Gestalt soit actuellement désuet (on préfère "structure", qui a une connotation plus mathématique) ou galvaudé (voir certaines doctrines de para-psychologie), les observations auxquelles elle a donné lieu dans le domaine de la perception ne sont pas contestées. Elle explique par exemple les problèmes de constance (de taille, de teinte, de forme, dans le domaine visuel) par la persistance de certains rapports internes d'une figure, assurant son unité perceptive. D'une manière générale elle admet qu'une Forme peut subir certaines transpositions, distorsions, altérations, sans perdre son caractère unitaire.

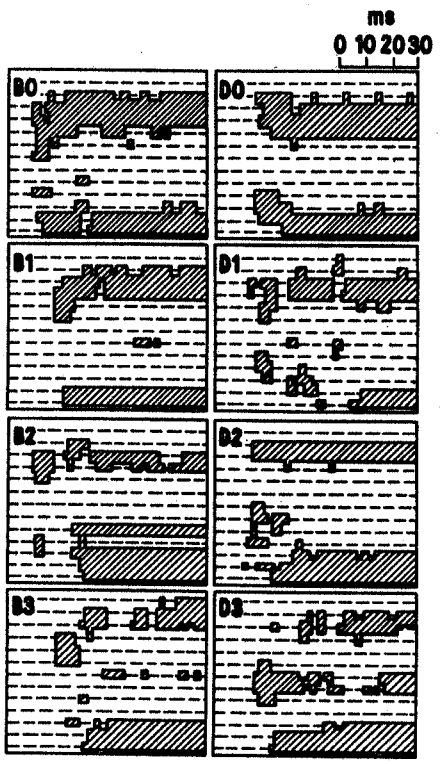
Malheureusement la théorie ne se prête pas facilement à une formalisation mathématique, et reste relativement peu prolixe sur la question de l'apprentissage. Ces aspects négatifs ne nous semblent pas suffisants pour rejeter en bloc toutes les observations faites ci-dessus.

IV. UN EXEMPLE

La figure suivante représente le début des lettres B et D prononcées par des locuteurs américains (2 locuteurs 0 et 1, 2 locuteurs 2 et 3) (6). Ces séquences, prononcées isolément, sont enregistrées via une ligne téléphonique et analysées au moyen d'un banc de 13 filtres approximativement répartis selon une échelle de Bark entre 300 et 3000 Hz. La représentation proposée est obtenue après un traitement spectral (voisin d'une double dérivation) mettant en évidence des maximums spectraux que l'on se gardera bien d'appeler formants.

Les phonèmes B et D en position initiale sont extrêmement voisins ; ils ne se différencient phonologiquement que par le trait grave/aigu. Comme il s'agit de parole téléphonique, les fréquences élevées caractérisant éventuellement l'explosion ont disparu. Ces mots sont néanmoins intelligibles avec un faible taux d'erreur. Soumis à un processus de reconnaissance par mots isolés monolocuteur, le taux d'erreur apparaît également comme très faible, surtout pour les locuteurs

TRAITS SPECTRAUX GROSSIERS DES MOTS "B" et "D"



0 et 1. Il est de l'ordre de 20% pour les locuteurs masculins et de 40% pour les locuteurs féminins, l'ensemble de référence étant constitué des lettres P B T D V Z prononcées isolément par des locuteurs américains. Cependant quand on passe à une reconnaissance multi-locuteur, les taux d'erreur augmentent et se rapprochent de ceux que donnerait un choix aléatoire. Il nous faut expliquer toutes ces observations.

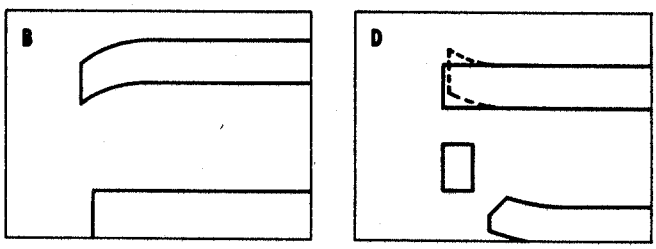
La bonne reconnaissance en mode mono-locuteur provient des différences spectrales observées sur la voyelle /i/ après stabilisation. En éliminant complètement la partie transitoire initiale, il reste suffisamment d'information dans le spectre vocalique pour différencier B de D de manière fiable, vu les conditions bien contrôlées de l'élocution et de la prise de son.

Malheureusement, le changement spectral quand on passe de B à D n'est pas le même pour tous les locuteurs. Même avec la faible précision de notre représentation, il est évident que le locuteur 2 ne réalise pas la différence entre B et D de la même manière que les autres, puisque la bande fréquentielle supérieure monte au lieu de descendre. L'indice acoustique utilisé par la reconnaissance monolocuteur (spectre de la voyelle stable) n'avait donc pas de valeur universelle (i.e. phonétique).

Maintenant, comment fait l'auditeur humain pour reconnaître ces mots avec une quasi-certitude . Il n'y a pas d'indice bien net, que l'on puisse retrouver

dans les quatre réalisations de chaque mot et qui soit caractéristique de l'opposition B/D. Bien sûr, la transition montante de la bande supérieure apparaît souvent pour B ; on peut discerner une évolution descendante de la bande inférieure pour D ; le calage temporel des deux bandes joue peut-être un rôle, mais aucun indice n'apparaît comme universel et décisif ; les bandes elles-mêmes ne peuvent pas toujours être identifiées avec certitude.

Alors il nous faut penser à la configuration d'ensemble, c'est-à-dire à la Gestalt. Si nous essayons de discerner un invariant dans les représentations de B et D, nous aboutissons aux schémas suivants, qui pourraient représenter les "bonnes formes" associées à B et D :



Naturellement il ne s'agit ici que d'un exemple, et nous avons pleinement conscience des réserves à faire, car l'analyse effectuée n'est pas forcément conforme à celle que fait l'oreille, et la transposition sur le plan visuel peut être génératrice d'erreurs de jugement. Mais nous souhaitons surtout promouvoir un cadre de pensée. Dans l'exemple précédent, nous avons le sentiment qu'il est vain de définir des indices acoustiques et même de chercher des relations entre eux. Chaque séquence est perçue d'emblée comme un tout, de même que la vue d'un visage et sa reconnaissance ne nécessitent pas sa décomposition en nez, yeux, bouche, oreilles, sauf si cela est requis par les niveaux supérieurs - et dans ce cas on ne "voit" pas, on scrute.

V. CONCLUSION

Nous avons voulu montrer plusieurs choses. Tout d'abord, ce qu'on appelle "variabilité" de la parole doit être précisé. Ce qui est variable, c'est l'aspect sous lequel se présente la réalisation du message quand on ne considère qu'un seul type d'information transmise et qu'on ne peut figer tous les autres. Si l'on arrive à fixer tous les aspects du message en même temps, alors il subsiste une variabilité intrinsèque, sans doute assez réduite, qui tient à l'incapacité de nos systèmes phonatoire et auditif à produire un signal physique invariable et à le percevoir de manière absolue. Le second point est que la variabilité au sens

large, c'est-à-dire l'ensemble des variations significatives du message, loin d'être un inconvénient, est peut être la clef qui nous permettra de comprendre les mécanismes perceptifs et cognitifs associés au décodage de la parole. Quant au troisième et dernier point, c'est la suggestion, que nous avons déjà émise maintes fois, qu'il nous faut aller plus loin que les démarches analytiques et associationnistes que nous suivons actuellement en reconnaissance automatique de la parole, et chercher à caractériser des entités globales comme celles que la thèse de la Forme a dégagées. Il nous faut d'autres outils de pensée, et d'autres machines, peut être des machines cellulaires à apprentissage sur lesquelles on peut émettre quelques hypothèses fonctionnelles ... mais ce sera pour une autre fois.

VI. BIBLIOGRAPHIE ULTRA-SOMMAIRE

1. JAKOBSON, R., FANT, G., HALLE, M. : Preliminaries to Speech Analysis : the distinctive features and their correlates - MIT Press, 1952.
2. GALF-GRECO 39 - Séminaire sur le Décodage Acoustico-Phonétique de la Parole Toulouse, 1981.
3. GUILLAUME, P. - La psychologie de la Forme - Flammarion, Paris, 1937.
4. LEIPP, E. - Information sémantique et parole : essai d'une Gestalt-Theorie - Bulletin du Groupe d'Acoustique Musicale n° 22, juin 1966.
5. LIENARD, J.S. - Les processus de la communication parlée - Masson, Paris, 1977.
6. LIENARD, J.S., SOONG, F.K. : On the use of transients in Speech Recognition IEEE-ICASSP, San-Diego, 1984.

FENETRE DE PRELEVEMENT TEMPOREL DES INDICES D'OCCLUSIVES

W. SERNICLAES : Institut de Phonétique de l'Université Libre de Bruxelles.

---

1. Introduction

Les recherches sur l'identification des occlusives suggèrent que la détente de l'occlusion contient de l'information invariante. Le prélèvement des 40 ms initiales de syllabes CV, prononcées isolément, fournit des indices suffisants pour la perception du lieu d'articulation et du voisement de l'occlusive (Tekieli & Cullinan, 1979).

L'appariement visuel du spectre statique prélivé dans les 20 ou 30 ms initiales de la détente avec des cibles spectrales correspondant aux lieux d'articulation permet d'atteindre des scores de reconnaissance proches de 85% (Blumstein & Stevens, 1978). La classification visuelle des spectres dynamiques prélevés durant les 40 ms initiales de la syllabe aboutit à des scores semblables (88%) (Kewley - Port, 1983). Dans chacune de ces deux approches la procédure d'analyse est fixe, sans ajustements contextuels. Cet aspect de la procédure pourrait être à l'origine de la variabilité des scores d'identification. Pour le voisement, le score de reconnaissance automatique passe de 97 à 99% lorsque l'on tient compte des effets des lieux d'articulation sur le VOT (Edwards, 1978). La nécessité des réajustements contextuels devient d'ailleurs évidente lorsque les syllabes sont prononcées en contexte de phrase (Amerman & Parnell, 1984).

En français, la perception du voisement des occlusives dépend simultanément de la présence/absence de voix durant l'occlusion et des caractéristiques de la détente (Serniclaes & Bejster, 1979; Serniclaes, 1979). Les données acoustiques présentées ici montrent que la présence/absence de périodicité au voisinage de la détente de l'occlusion permet d'opérer une séparation fiable entre les occlusives /ptk/ et /bdg/ en position intervocalique. Le prélèvement temporel de cet indice doit cependant être ajusté en fonction de la vitesse intrasyllabique pour tenir compte des effets des lieux d'articulation sur le VOT (Lisker & Abramson, 1967) et des effets de l'accent sur la durée de l'occlusion sourde (Wajskop, 1979).

Les données ont été receuillies dans 2 corpus, l'un constitué de 90 syllabes CV prononcées isolément, l'autre de 870 logatomes VCV prononcés dans différentes conditions. Parmi les divers indices temporels et spectraux qui

ont été mesurés, seuls le VOT (en CV et VCV), la durée du bruit de friction (en CV) et la durée de l'intervalle de silence qui précède la détente (IS) seront examinés ici. Les mesures de durées segmentales ont été publiées antérieurement (Wajskop, 1979). L'ensemble des résultats fera l'objet d'une publication ultérieure (Serniclaes, en prép.)

2. Procédure

2.1 Cadre CV

Locuteurs: 5 francophones belges masculins

Corpus: 18 syllabes CV correspondant aux 6 occlusives orales /ptk, bdg/ suivies des voyelles /a, u, i/.

Enregistrement: en chambre sourde, par micro Neuman U-67 et enregistreur Ampex-300 à 38 cm/s. Une erreur de prononciation a été décelée lors du contrôle auditif, ce qui nous a amenés à reprendre la procédure d'enregistrement pour l'un des locuteurs.

Mesures: VOT et durée du bruit de friction, sur spectrogrammes à bande large (300 Hz - Voiceprint 700). Le VOT a été mesuré en prennant l'intervalle de temps entre le début des vibrations périodiques et le transitoire de détente. Le bruit de friction a été mesuré à partir de la détente (inclue). Le repérage visuel de la fin de la friction était basée sur 2 critères: arrêt, ou chute de niveau, dans les formants supérieurs et début des transitions. Les mesures relatives aux occlusives sourdes sont incomplètes. Ceci vient du fait que la distinction entre friction et aspiration s'est avérée difficile a établir pour les syllabes /pa, pi, pu, tu/.

2.2 Cadre VCV

Locuteurs: 10 francophones masculins, âgés de 20 à 30 ans, originaires de la région bruxelloise et de familles appartenant à la classe moyenne.

Corpus: 87 logatomes VCV correspondant aux 6 occlusives orales dans 3 contextes vocaliques symétriques /a, u, i/ et prononcés dans 5 conditions différentes: (1) isolément; (2) accentués en séquence "dites VCV"; (3) inaccentués en séquence "dites VCV pour moi"; (4) accentués dans un mot en fin de phrase (sauf pour /upu, ubu, ugu/ et ce faute de mots adéquats); (5) inaccentués en phrase.

Enregistrement: dans des conditions semblables à celles du corpus CV.

Mesures: VOT et intervalle de silence pré-détente (IS), sur oscillogrammes (oscillomink Siemens) à 200 mm/s et avec contrôles sur sonogrammes. Les mesures d'IS n'ont été prises que dans les séquences insérées en mots, accentués et inaccentués.

3. Résultats

3.1 Cadre CV

Le VOT négatif des /bdg/, est de -130 ms en moyenne (σ=26; Fig. 1). Le prévoisement devient plus bref avec la rétraction du lieu d'articulation de l'occlusive (S à P<. 005). Ni l'effet du lieu de la voyelle, ni l'interaction occlusive - voyelle (Fig. 1) ne sont significatifs.

Le VOT positif des /ptk/ est de 31 ms en moyenne (σ=15; Fig. 1). Il s'allonge avec la rétraction du lieu del'occlusive (p<.005), avec la fermeture de la voyelle (p<. 01). Les effets de l'occlusive et de la voyelle sont interactifs (p<. 05). Bien que le VOT des dentales soit généralement plus long que celui des labiales, la tendance s'inverse devant /u/ (Fig. 1). L'effet de la fermeture vocalique ne se retrouve pas pour /pi/. En ce qui concerne les /bdg/, nous avons trouvé quelques cas d'interruption de voix (3/45) au voisinage de la détente,mais ces arrêts ne dépassent jamais deux périodes de fondamentale (~20 ms). L'erreur de prononciation dont nous avons fait mention dans la procédure provient d'une interruption de voix de 31 ms durant la friction d'une occlusive dorsale.

Bien que la durée de la friction des occlusives sourdes ne soit pas toujours mesurable, l'examen de la Fig. 2 suggère que le bruit de friction n'est pas systématiquement plus long pour les sourdes. Selon le contexte la friction est tantôt plus longue pour les sourdes, tantôt plus brève mais les écarts sont faibles et s'annulent dans l'ensemble. La durée moyenne est de 17 ms (σ =11), soit 14 ms de moins que celle du VOT positif. Les variations contextuelles sont similaires. La friction s'allonge généralement avec la rétraction du lieu de l'occlusive (p<.005) et avec la fermeture de la voyelle (p<. 01). L'interaction occlusive-voyelle (Fig. 2) n'est cependant pas significative.

Les variations contextuelles du VOT et de la durée de la friction proviennent de contraintes aérodynamiques liées aux changements de vitesse intrasyllabique (Haag, 1979; Serniclaes, en prép.). La durée de la friction des voisées ne fournit que des indications très grossières sur la vitesse des transitions des sourdes homorganiques prononcées pour le même locuteur. Néanmoins, la corrélation entre le VOT des sourdes et la friction des voisées correspondantes est relativement forte (r_{BP}=0.62; 38% de variance expliquée; S à p<. 001). En gros, le VOT recouvre l'intervalle de friction et le début des transitions formantiques. Le délai entre la fin de la friction et le début de la voix, ou intervalle d'aspiration (~ 15 ms en moyenne), n'est cependant pas constant; il s'allonge lorsque les transitions sont plus lentes.

3.2 Cadre VCV

Les distributions des VOT des /ptk/ intervocaliques, isolés ou accentués, sont quasi-identiques à celle des prévocaliques. Dans chaque cas le VOT moyen est légèrement supérieur à 30 ms et l'écart-type est de ~15 ms (Tab. 1). Les VOTs des occlusives inaccentuées (logatomes ou mots en VCV) sont légèrement plus brefs, de 5 ms en moyenne (S à $p < .001$). La réduction n'est cependant pas uniforme. Les écart-types diminuent, ce qui signifie que la réduction est plus faible pour les VOTs brefs. De même, la variabilité des valeurs inférieures à la moyenne est plus faible au sein de chaque contexte, ainsi que l'indiquent les coefficients de symétrie (q1; Tab. 1).

Cadre de prononciation	$\bar{m}$	$\bar{\sigma}$	$\bar{q_1}$	N
CV, logatomes isolés	31	15	0,64	45
VCV, logatomes isolés	31	15	0,54	90
VCV, logatomes en contexte accentué	32	14	0,63	90
VCV, logatomes en contexte non-accentué	26	11	0,81	90
VCV, mots en contexte accentué	32	14	0,81	80
VCV, mots en contexte non-accentué	29	13	0,55	90

Tab. 1

Caractéristiques des distributions des VOTs (ms) des occlusives sourdes (moyennes, écart-types, coefficients de symétrie, nombres d'observations).

Les valeurs minimales du VOT sont pratiquement toujours supérieures (92%) ou égales (8%) à 10 ms, et ce même lorsque l'occlusive n'est pas accentuée. Le VOT est beaucoup moins stable en anglais en raison des différences allophoniques entre les positions CV et VCV et d'une influence beaucoup plus nette de l'accent (Flege & Brown, 1982; Lisker & Abramson, 1967).

La Fig. 3 donne les distributions du VOT et de l'IS pour les /ptk/ intervocaliques insérés en mots. On voit que la durée de l'IS dépend fortement de l'accent. De 84 ms, en moyenne (σ =25), dans les mots accentués l'IS se réduit a 50 ms (σ =20) dans les mots inaccentués. La réduction de l'IS est liée à celle de la durée de l'occlusion (Wajskop, 1979). En français, les mouvements d'ouverture glottale et d'occlusion orale se développent de concert (Benguerel et al., 1978). En anglais, le mouvement de la glotte est décalé et son ouverture est maximale lors de la détente de l'occlusion (Löfquist, 1980). Les effets de l'accent sur l'IS et le VOT semblent donc provenir de la glotte. L'accent se manifeste d'avantage aux moments où la glotte est ouverte: durant l'occlusion

en français et lors de la détente en anglais.

L'IS est relativement fiable en dépit de sa variabilité. Pour les sourdes il dépasse généralement 30 ms (92% des cas) et on ne relève que 4% de sonorisations complètes de l'occlusion (IS=0, Fig. 3). Nous n'avons pas constaté d'interruptions de voix pour les voisées intervocaliques. Les vibrations périodiques sont toujours très apparentes sur les oscillogrammes. L'examen des sonagrammes fait apparaître des chutes d'intensité lors de la détente, sans arrêts de périodicité. En position prévocalique, les arrêts de voix proviennent d'imprécisions dans le contrôle des relations temporelles entre articulateurs (Serniclaes et al., en prép.). Ce problème ne se pose pas lorsque le début de l'activité laryngée précède l'occlusion, comme c'est le cas en VCV.

4. Discussion

4.1 Prélèvement de la périodicité

Si la localisation de la fenêtre de prélèvement de la périodicité (FP) est fixe, les erreurs d'identification seront inévitables. Le taux d'erreurs peut cependant être maintenu en dessous de 10% en prennant un intervalle qui va de 30 ms avant la détente à 10 ms après la détente (Fig. 3), soit 40 ms, ce qui est suffisant pour détecter la périodicité (Doughty & Garner, 1947; Rietveld, ce volume). Ce sont les variations complémentaires de la durée de l'occlusion et du VOT (Serniclaes & Bejster, 1974) qui permettent de contenir le taux d'erreurs. A une exception près les occlusives sourdes dont l'IS est inférieur à 35 ms présentent un VOT supérieur à 10 ms, et vice-versa.

L'ajustement de la FP en fonction de la vitesse intrasyllabique permet d'accroître la fiabilité. La procédure d'ajustement la plus adéquate consiste à déplacer la fermeture de la FP en fonction de la vitesse des transitions post-consonnantiques. La Fig. 4 montre que l'IS ne dépend pas systématiquement des lieux d'articulation, surtout lorsque l'occlusive n'est pas accentuée. Il est donc préférable d'ajuster la FP en fonction de la vitesse des transitions -CV, dont dépend le VOT (§ 3.1). Dans cette optique, l'IS fournit un indice complémentaire dans les contextes où le VOT est relativement bref.

Le cumul de l'IS et du VOT donne la durée de l'arrêt de voix. Les arrêts de voix sont, en moyenne, de 116 ms (σ=22) pour les /ptk/ accentués et de 84 ms (σ=20) pour les non-accentués. Dans l'ensemble, la durée moyenne de l'arrêt de voix des /ptk/ intervocaliques de l'ordre du dixième de seconde et il est pratiquement toujours (169 cas sur 170) supérieur à 35 ms. Une localisation flexible de la FP ouvre donc la possibilité d'identifier le voisement de l'occlusive intervocalique avec des taux d'erreurs inférieurs à 1%.

Pour les occlusives prévocaliques, les risques d'erreur proviennent des cas de désonorisation de l'occlusion des /bdg/ (Durand, 1956). La désonorisation dépend du dialecte et de facteurs socio-culturels (Goudaillier, 1983). Les pourcentages de désonorisation des voisées initiales augmentent avec la rétraction du lieu d'articulation de l'occlusive (Goudaillier,1983). De nouveau, la localisation de l'intervalle de temps critique, durant lequel la fiabilité de la présence/absence de voix est optimale, dépend de la vitesse des transitions.

4.2 Autres indices

Les variations contextuelles du VOT s'accompagnent de différences qualificatives. Plus les transitions intrasyllabiques sont lentes, plus le bruit de friction est long et plus la distance spectrale entre l'occlusive et la voyelle est faible (Serniclaes, en prép.). Le ralentissement des transitions renforce la distinctivité des indices spectraux fournis par la détente de l'occlusion (Serniclaes, 1979). Comme pour le lieu d'articulation (Dorman et al., 1977), on se trouve en présence de variations complémentaires entre indices dynamiques (VOT et transitions) et statiques (force et spectre du bruit d'explosion). Ceci suggère que l'ajustement contextuel de la fenêtre de prélèvement doit aller de pair avec une modification des processus d'analyse.

Références

Amerman, J.D. and Parnell, M.M.(1984) "Variable perceptual potency of the initial 20 ms time Window". J. of Phonetics 12, 1-7.

Benguerel, A.P., Hirose, H., Sawashima, M. and Ushijima, T. (1978) "Laryngeal control in French stop production: a Fiberscopic, Acoustic and Electromyographic study". Folia Phoniat. 30, 175-198.

Blumstein, S.E. and Stevens, K.N. (1979) "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants" J. Acoust. Soc. Am. 66, 1001-1017.

Durand, M. (1956) "De la perception des consonnes occlusives: questions de sonorité". Word 12, 15-34.

Dorman, M.F., Studdert-Kennedy, M. and Raphaël, L.J. (1977) "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues" Perception & Psychophysics 22, 109-122.

Doughty, J.M. and Garner, W.R. (1947) "Pitch characteristics of short tones". J. Exp. Psychol. 37, 351-365.

Edwards, T.J. (1978) "A probabilistic vector model for identification of intervocalic stop consonants" IEEE Int. Conf. on Acoustics, Speech and Signal Processing (April 1978).

Flege, J.E. and Brown, W.S. (1982) "The voicing contrast between /p/ and /b/ as a function of stress and position-in-utterance". J. of Phonetics 10, 335-345.

Goudaillier, J.P. (1983) "Diverses possibilités de matérialisation du trait de voisement". Revue d'Acoustique, Hors Série 11e I CA, Vol. 4, 267-270.

Haag, W.K. (1979) "An articulatory experiment on Voice Onset Time in German stop consonants". Phonetica 36, 169-181.

Kewley - Port, D. (1983)"Time-varying features as correlates of place of articulation in stop consonants". J. Acoust. Soc. Am. 73, 322-335.

Klatt, D.H. (1975) "Voice onset time, frication and aspiration in word initial consonant clusters". J. Speech & Hearing Res. 18, 686-705.

Lisker, L. and Abramson, A.S. (1967) "Some effects of context on Voice Onset Time in English stops". Language and Speech 10, 1-28.

Löfqvist, A. (1980) "Interarticulator programming in stop production". J. of Phonetics 8, 475-490.

Rietveld, A. (ce volume) "Evaluation perceptive de la détection du voisement".

Serniclaes, W. (1979) "Sur la dissociation entre périodicité, bruit et fréquence fondamentale en tant qu'indices de voisement des occlusives du français" Rapport d'Activités de l'Institut de Phonétique de Bruxelles 13, 71-93.

Serniclaes, W. (en préparation) "Acoustic description of voicing contrasts in French stops".

Serniclaes, W. et Bejster, P. (1974) "Influence du contexte vocalique sur la perception du voisement des occlusives". Actes des 5e J.E.P. (Galf), Orsay, 10-18.

Serniclaes, W. and Bejster, P. (1979) "Cross-language differences in the perceptual use of voicing cues". Amsterdam Studies in the Theory and History of Linguistic Science IV, Vol. 9, H. & P. Hollien Eds.; J. Benjamins; 755-764.

Serniclaes, W., D'Alimonte, G. and Alegria, J. (en préparation) "Production and perception of stop consonants by the French speaking deaf".

Tekieli, M.E. and Cullinan, W.L. (1979) "The perception of temporally segmented vowels and consonant-vowel syllables". J. Speech & Hearing Res. 22, 103-121.

Wajskop, M. (1979) "Segmental durations of French intervocalic plosives". Frontiers of Speech Communication Research, B. Lindblom and S. Öhman eds.; Academic Press; 109-123.

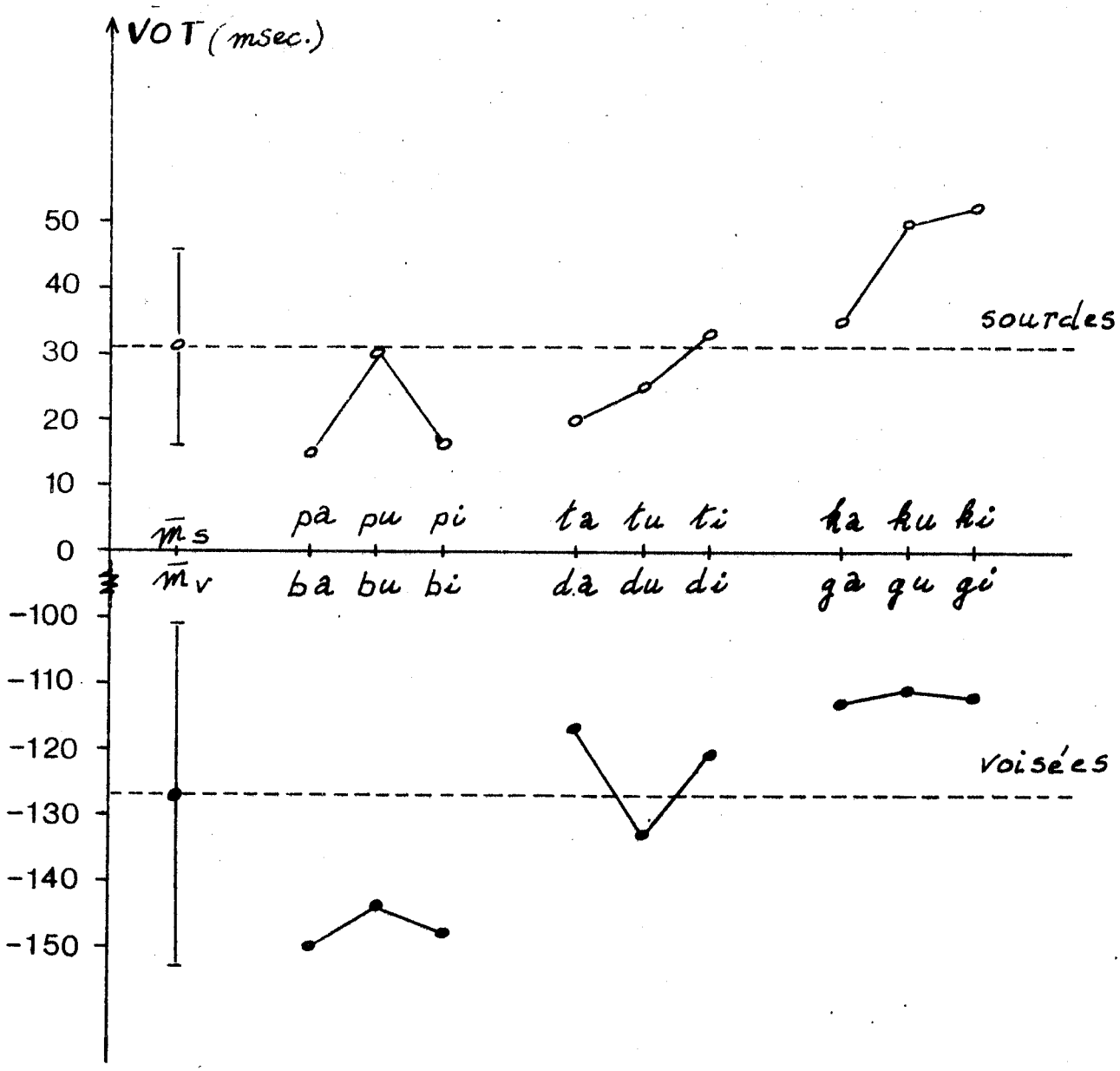


Fig. 1. VOT des occlusives prévocaliques.

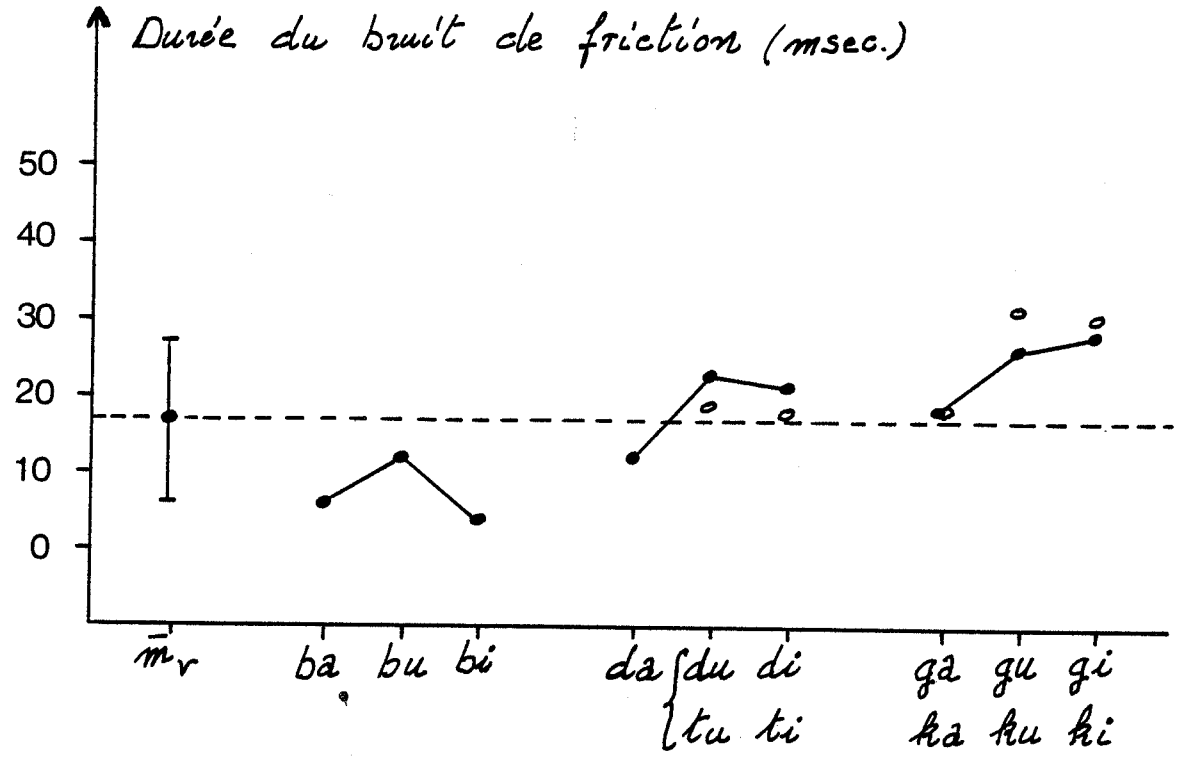


Fig. 2. Durée du bruit de friction des occlusives prévocaliques. Les mesures relatives aux occlusives sourdes sont incomplètes (voir texte).

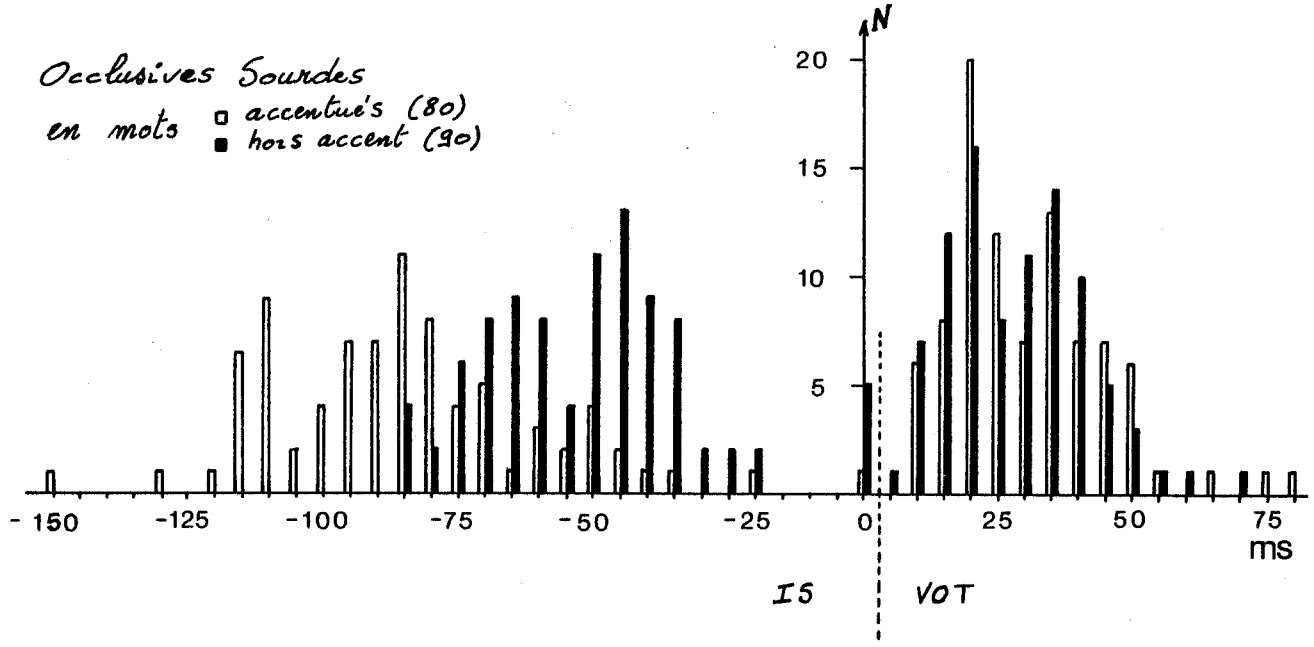


Fig. 3. Distributions de l'IS et du VOT des occlusives intervocaliques insérées en mots.

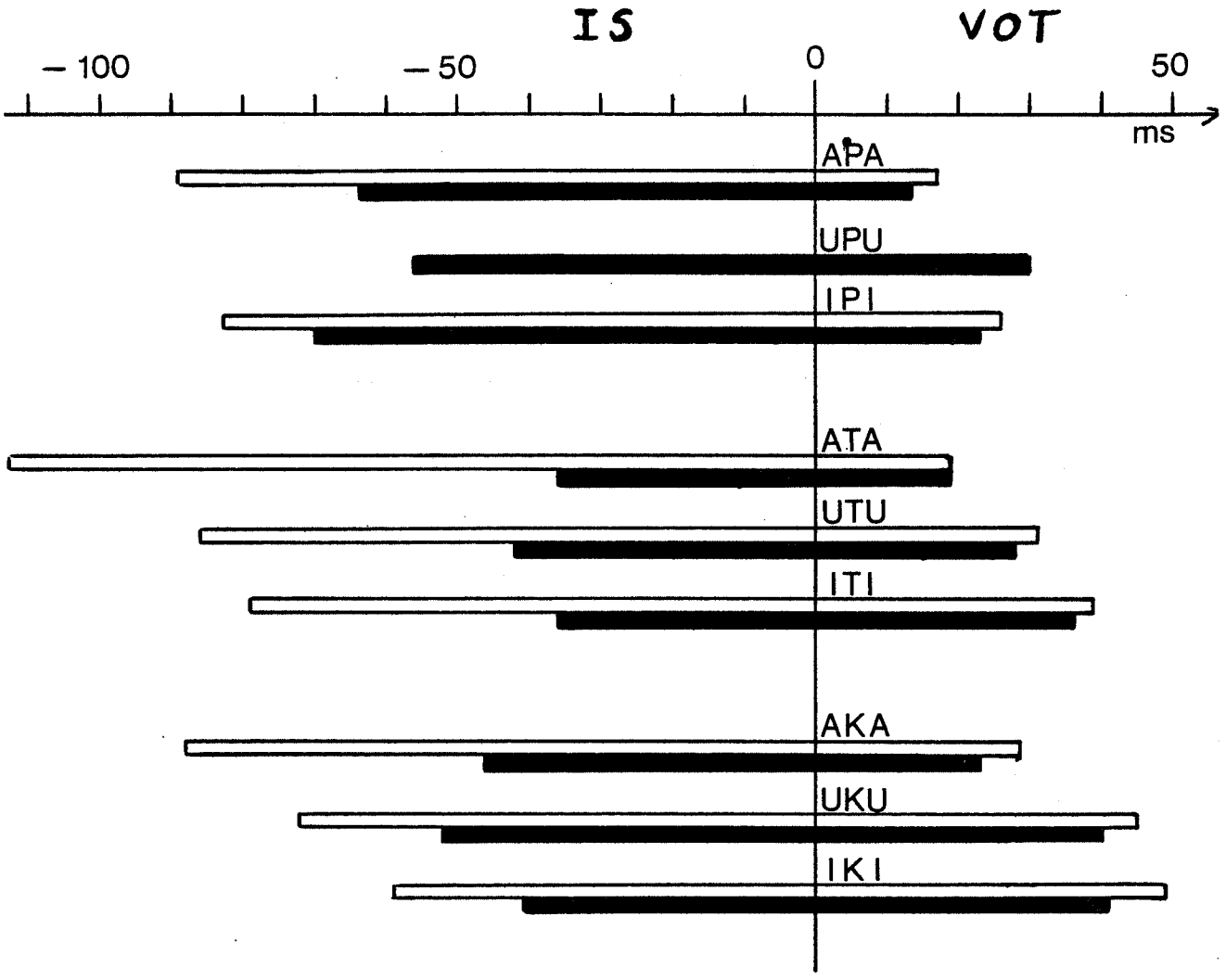


Fig. 4. Valeurs moyennes de l'IS et du VOT pour chaque contexte.

SYNTHESE MULTILOCUTEUR DE HAUTE QUALITE

C. BIETRY : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France.

I. INTRODUCTION

L'objet de notre recherche est d'obtenir, par une synthèse de très grande qualité, la connaissance et la maîtrise du système de paramètres caractéristiques des voix de différents locuteurs, et ce en vue de formaliser d'éventuelles règles de passage d'un locuteur à un autre, pour un même support phonétique. C'est une recherche d'ordre fondamental qui, dans son aspect multivoix est semble-t-il tout-à-fait originale. Depuis les années 50 et les premiers travaux de DELATTRE sur le Pattern-Play-Back, on ne compte plus les travaux utilisant la synthèse pour mettre à jour les caractéristiques acoustiques et phonologiques de la parole. Par ailleurs, on a aussi beaucoup étudié par analyse acoustique ou articulatoire, tout ce qui dans la production phonétique ressortissait de la variabilité inter-locuteur : dispersion des plages formantiques, réalisations prosodiques, etc. Enfin plus récemment, les impératifs de la reconnaissance automatique de la parole et l'application à la reconnaissance du locuteur ont amené les chercheurs de ce domaine à mettre au point des méthodes de traitement des indices acoustiques dans un contexte multilocuteur, mais méthodes en général aveugles à ce qui peut faire la spécificité paramétrique d'une voix par rapport à une autre.

Par contre, à notre connaissance, aucune recherche à ce jour n'a porté sur l'étude fine des caractéristiques paramétriques de différentes voix, utilisant une démarche d'analyse-synthèse qui permet de tester perceptivement les hypothèses concernant l'identité d'une voix. Une raison en est sans doute que si on sait depuis longtemps reproduire par synthèse des voix intelligibles mais à sonorité toujours artificielle, ce n'est que depuis peu que le perfectionnement des outils de synthèse permet la qualité "voix naturelle".

Il s'agit donc pour nous, par une démarche d'analyse synthèse se situant sur une base acoustique et non articulatoire, de mettre en évidence les facteurs acoustiques de l'intelligibilité, du timbre, de l'accent, du type de voix ... etc pour plusieurs locuteurs et d'établir quelles relations entre ces facteurs sont caractéristiques d'un locuteur donné. Nous signalons, que dans le même temps, une étude est menée similairement en collaboration avec Mr SANTERRE de l' Université de Montreal, sur les caractéristiques de deux réalisations dialectales d'une même langue : le français standard et le français quebecois, en vue de formaliser les

règles de passage d'un dialecte à un autre. Ces deux études utilisent la même méthode et posent des problèmes tout-à-fait convergents.

II. METHODE D'ANALYSE-SYNTHESE

Nous avons délibérément choisi une méthode d'analyse-synthèse non automatique : les paramètres nécessaires à la synthèse sont extraits à la main ... et à l'oeil, principalement à partir d'une analyse numérique par banc de filtres. Etant donné notre formation de phonéticienne et notre volonté de maîtriser ce qui, dans un système paramétrique est nécessaire à la reproduction du timbre et du naturel, nous avons préféré avoir un contrôle permanent sur les valeurs extraites de l'analyse et introduites en synthèse. La méthode est coûteuse en temps de préparation, mais nous a permis d'acquérir une grande maîtrise du processus, ce qui nous permet aujourd'hui d'obtenir dès les premiers essais des synthèses de bonne qualité.

1°) L'analyse

Nous nous servons de deux logiciels d'analyse par banc de filtres du 2^{e} ordre, délivrant sur 28 ou 32 canaux une quantification non normalisée de l'énergie toutes les 10 ms, sur l'intervalle 0-5000 Hz. Le premier logiciel reproduit une analyse de type sonagraphe avec 32 filtres de 300 Hz de bande passante et un recouvrement de 150 Hz. Le deuxième utilise une répartition de 28 filtres suivant une échelle de Barks, ce qui donne une meilleure précision sur les basses fréquences. L'intérêt de ces analyses est de fournir sur un même support une image spectrale précise dans les trois dimensions : temps, fréquence et intensité.

Le principe consiste à comparer les spectres numériques et à chercher à obtenir le meilleur appariement possible à chaque instant et dans chaque canal (et non seulement dans les zones portant l'information formantique ou consonantique). L'appariement numérique parfait étant impossible à atteindre le résultat est jugé satisfaisant lorsqu'à partir d'un bon appariement spectral la comparaison auditive entre l'original et la synthèse est bonne. Nous nous servons pour ça de tests d'audition.

2°) La synthèse

Nous avons opté pour une synthèse à formants parallèles, seule capable à notre avis de nous fournir avec qualité une bonne reproduction de plusieurs types de voix. Après divers essais, nous avons retenu le synthétiseur numérique de Klatt que nous n'utilisons que dans sa version "tout parallèle". Bien qu'imparfait à divers titres, entre autre parce que non optimisé pour le fonctionnement "tout parallèle", le synthétiseur de Klatt présente cependant un système de paramétrisation très complet et très souple, et est implanté dans plusieurs laboratoires en France et à l'étranger.

Nos synthèses ont toutes été obtenues en jouant sur une vingtaine de paramètres :

- fréquence (Fi), intensité (Ai) et bande passante (Bi) de 4 filtres pour les sons voisés - fréquence fondamentale (Fø), amplitude de voisement (Av) et un facteur de pente spectrale de l'onde globale - et facultativement 3 paramètres pour le formant de nasalisation (de bonnes nasales peuvent être réalisées sans) et 1 filtre supplémentaire pour certaines fricatives.

Chaque paramètre étant modifiable au mieux toutes les 10 ms.

III. RESULTATS

1) Remarques d'ensemble

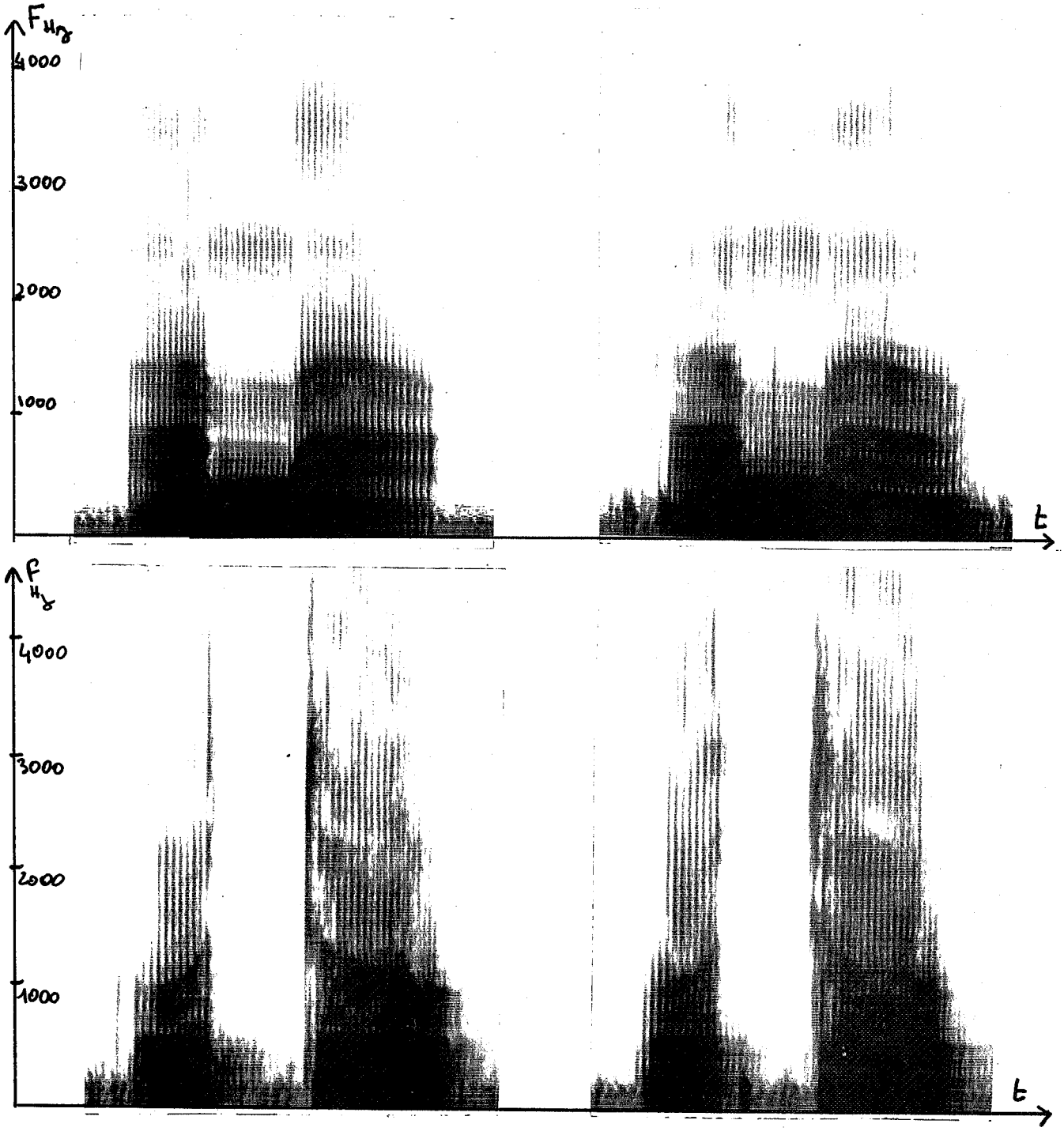
Au cours de nos différents travaux, nous avons synthétisé les voix d'une dizaine de locuteurs et leur reproduction a montré que dans la plupart des cas, on peut obtenir avec notre méthode ... et beaucoup de temps, des synthèses très proches perceptivement des voix originales, présentant toutes une très grande qualité de naturel et permettant en particulier de reconnaître le locuteur.

Dans quelques cas cependant, et malgré de très nombreux essais, il subsiste des différences de qualité qui vraisemblablement sont à imputer au modèle de source du synthétiseur. En effet, dans le système de Klatt, la source est le produit d'une série d'impulsions, filtrées par un filtre passe-bas, et le résultat est un spectre d'onde globale dont les caractéristiques sont assez peu modifiables. Dans ces conditions, certaines voix resteront toujours plus difficiles à imiter que d'autres, et cet écueil dans nos travaux apporte une contribution supplémentaire à la mise en évidence de l'importance de la forme d'onde globale dans la reproduction de la parole. En l'occurrence, il nous a été plus facile d'imiter les voix au fondamental grave et au timbre riche.

Par ailleurs, la réalisation de tests permettant de contrôler la qualité de nos synthèses n'est pas sans poser de nombreux problèmes, et la mise au point de tests adéquats est encore à faire dans une large mesure. En effet il nous faut évaluer nos résultats définitifs en intermédiaires suivants plusieurs critères :

1- qualité de la reproduction au niveau phonétique (intelligibilité)
2- qualité de naturel
3- et surtout proximité du timbre du locuteur (ou distance lorsque nous déformons sciemment nos synthèses pour passer d'un locuteur à l'autre).

Si le premier critère ne pose pas de problème particulier, les deux autres dont l'importance est évidente n'en sont pas moins délicats à définir et difficiles à cerner au moyen de tests de perception quantitatifs.Le problème de l'identification absolue du locuteur à partir de la voix de synthèse est encore plus complexe : si au laboratoire, dans des conditions d'écoute spontanée de nos syn-



SYNTHESE MULTI- LOCUTEUR

Les mots ama et oto prononçés par 2 locuteurs sont représentés:

Où est l'original , où est la synthèse...?

fig. 0

thèses, nos collègues, en passant, identifiaient souvent sans peine le locuteur d'origine, même sur des mots très brefs, parce qu'ils cotoient tous les jours les personnes ayant servi de modèle, une quantification sérieuse de ce genre d'épreuve est difficile à conceptualiser. Nous poursuivons notre travail de réflexion sur ces points.

2) Résultats en synthèse multivoix

Dans une première étape, nous avons voulu vérifier si nos outils nous permettaient d'obtenir effectivement des synthèses quasi-indiscernables des voix d'origine. Nous avons donc travaillé sur deux mots brefs (ama et oto) prononcés chacun par six locuteurs masculins différents. Ces douze synthèses, de qualité inégale pour les raisons exposées ci-dessus, ont été présentées par paires (original/synthèse ou synthèse/original) à une vingtaine d'auditeurs dans un test d'évaluation de distance sur quatre niveaux. (voir figure 0)

Les résultats de ce test apportent plusieurs indications :

1- 9 des 12 synthèses sont estimées ayant une distance très légère, les 3 autres ayant une distance moyenne. Ceci nous a donc confirmé que notre méthode permettait d'obtenir des synthèses de très bonne qualité, sinon indiscernables de l'original.

2- La dispersion des réponses d'un auditeur à l'autre est beaucoup plus grande que ce à quoi nous nous attendions : même pour les synthèses qui font le meilleur score de ressemblance, il y a autant d'auditeurs (30%) pour juger leurs distances au modèle nulle qu'il y en a qui l'estiment importante. Il semble donc qu'il y ait une très grande subjectivité dans l'estimation de la proximité au locuteur.

3- Cette subjectivité s'appuie en partie sur les caractéristiques de la voix elle-même : la synthèse d'une voix "typée", même moins bonne sur le plan acoustique, sera jugée plus ressemblante au modèle que celle d'une voix neutre.

Il nous reste sans aucun doute à améliorer nos outils de travail, et plus particulièrement nos outils d'analyse et de comparaison spectrale des résultats.
Mais d'ores et déjà nous pouvons considérer que le système de paramètres définis pour chaque synthèse est une bonne image du support phonétique et des caractéristiques du locuteur.

3) Etude comparative des systèmes de paramétrisation

Nous n'hésiterons pas à dire qu'une telle étude, prise globalement, est au premier abord un véritable casse-tête chinois. En effet non seulement il y a variabilité interlocuteur pour chaque paramètre, mais, entre deux locuteurs, cette variabilité ne va pas dans le même sens pour tous les paramètres. Autrement dit il est difficile de dégager des relations de type homothétique ou "anamorphose" permettant de passer d'un locuteur à l'autre.

J.S. LIENARD a mis en évidence de telles relations entre des voix d'homme, d'enfant et de femme pour les paramètres hauteur moyenne et échelle formantique, relations caractéristiques justement du sexe et de l'âge du locuteur.

Mais pour un ensemble de locuteurs tous masculins, d'âge proche, les caractéristiques de timbre sont portées par des relations beaucoup plus complexes : notre locuteur 1 au fondamental moyen parmi les plus élevés (150 Hz) a un système formantique (F_1 F_2) plus grave que celui du locuteur 4 dont le fondamental moyen est le plus grave (110 Hz). D'une façon générale, pour ces six locuteurs, la dispersion de chaque paramètre par rapport à sa valeur moyenne n'excède pas $\pm$ 20%, ce qui au maximum, d'un locuteur à l'autre, donne une variabilité de l'ordre de 30 à 40%. Et c'est évidemment sur la fréquence fondamentale qu'on retrouve la plus grande variabilité.

Il nous a donc semblé nécessaire d'introduire des étapes dans notre recherche des relations inter-paramétriques caractéristiques d'une voix donnée : si de telles relations existent, et elles existent puisqu'elles sont "reconnues" à la perception (empreinte vocale d'un locuteur), elles sont vraisemblablement à aborder dans une démarche de type Gestalt. Or les outils méthodologiques d'une telle approche restent encore largement à définir, du moins pour ses applications scientifiques.

Nous nous sommes donc tournés vers une démarche plus classiquement analytique et, dans une deuxième étape actuellement en développement, nous testons la résistance de la qualité de nos synthèses en modifiant systématiquement les valeurs de certains paramètres. Une telle démarche devrait nous permettre d'établir une hiérarchie entre les paramètres quant à leur importance dans la reproduction de la spécificité d'une voix, et, à partir de là, d'aborder l'étude des relations caractéristiques de cette identité.

IV. INVARIANCE DU TIMBRE

Cette étude a été réalisée essentiellement à Montreal, dans le cadre de la collaboration dont il a été fait état plus haut. Sans exposer ici les tenants et aboutissants de cette recherche pluri-dialectale, nous indiquons cependant qu'elle nécessite, a contrario, d'éliminer de l'étude les variations inter-individuelles, pour ne mettre en évidence que les variations dialectales. Nous avons donc cherché des locuteurs capables de parler aussi bien les deux dialectes. C'est sur la voix d'une des personnes retenues que nous avons mené les expériences suivantes, car nous avions alors réalisé la synthèse d'une phrase relativement longue prononcée par ce locuteur. Notre propos, à Mr SANTERRE et à moi-même, était alors de faire parler quebecois cette personne, par synthèse, sans modifier ses caractéristiques individuelles, et les premiers résultats obtenus sont encou-

rageants.

1°) Essais sur l'intensité des formants

Une première série d'essais a porté sur l'intensité relative de 4 formants (Ai) et sur la bande passante du 1er formant (B1). Nous voulions tout d'abord vérifier si notre acharnement à vouloir définir ces paramètres avec précision afin d'obtenir le meilleur appariement spectral, était justifié sur le plan perceptif. Nous avons donc réalisé quatre nouvelles synthèses :

- synthèse n° 9 : A_3 et A_4 sont remplacés à chaque instant par $\frac{A_3+A_4}{2}$ (ce qui provoque des différences de + 6dB sur un formant à - 6dB sur l'autre)
- synthèse n° 8 : A_1 et A_2 sont remplacés à chaque instant par $\frac{A_1+A_2}{2}$ (A_3 et A_4 gardant les valeurs du n° 9)
- synthèse n° 7 : A_1, A_2, A_3 et A_4 ont la même valeur $\frac{A_1+A_2}{2}$ (ce qui en l'occurrence revenait à augmenter le poids des formants 3 et 4)
- synthèse n° 5 : toutes choses égales par ailleurs : doublement de B1.

Un test d'audition par paire avec estimation de distance nous a montré, a contrario, de ce que nous attendions, que ces modifications qui provoquent un désappariement spectral important, n'ont qu'un faible impact sur le plan perceptif : la 9, la 8 et la 5 sont jugées quasi-indiscernables d'avec la synthèse référence 1, la 7 a le plus mauvais score mais sa distance à la synthèse 1 ne la fait pas sortir du timbre du locuteur.

Les résultats ne sont paradoxaux qu'au premier abord et nous amènent plutôt à relativiser l'importance d'un appariement spectral parfait. Certes, si nous voulons obtenir des signaux synthétiques absolument indiscernables des signaux originaux, il faut l'appariement spectral le plus parfait possible. Mais si nous voulons reproduire une bonne intelligibilité phonétique, une bonne qualité de naturel, et une bonne proximité au timbre d'un locuteur, l'appariement spectral parfait n'est pas indispensable. La variabilité intra-individuelle elle-même en est la preuve.

En l'occurrence, pour nos synthèses 9 et 8, le poids respectif entre la partie haute et la partie basse du spectre est maintenue : on ne sort pas de "l'empreinte" du locuteur, à condition de respecter les caractéristiques d'attaque et de fin de voisement. Ces résultats nous ont amenés à chercher la limite de variabilité d'un paramètre à partir de laquelle on sort de la reconnaissance du timbre original. Nous avons mené ces expériences sur Fø, F4, et dans une moindre mesure A_4. Nous entendons les poursuivre sur d'autres paramètres.

2°) Essais sur la fréquence fondamentale

Le fréquence fondamentale est sans aucun doute un des paramètres les plus importants pour l'identité d'une voix, mais nous savons aussi qu'elle varie facilement d'un octave dans les productions d'un même locuteur. Il a aussi été établi

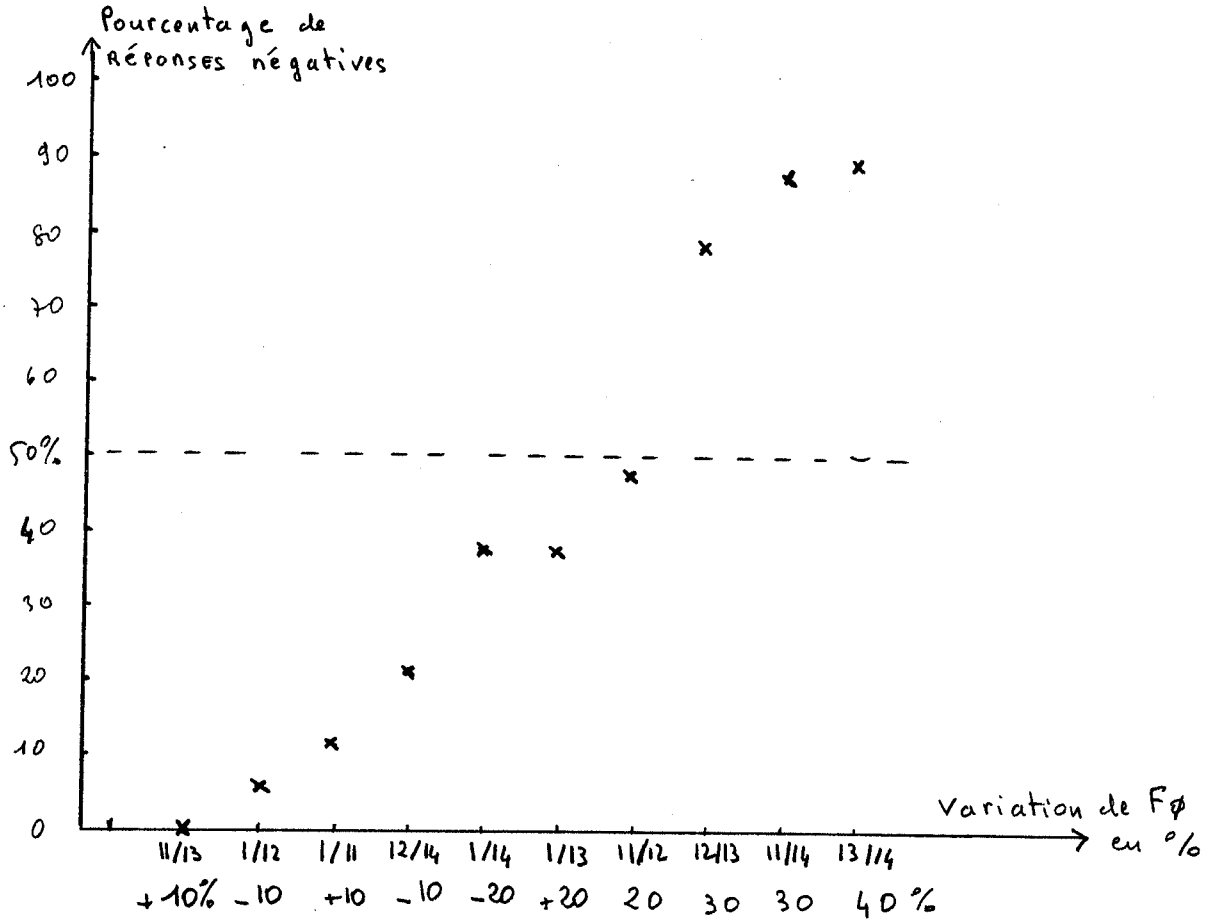


Fig 1 : PERTE DE L'IDENTITÉ VOCALE EN FONCTION DE LA VARIATION DE Fø

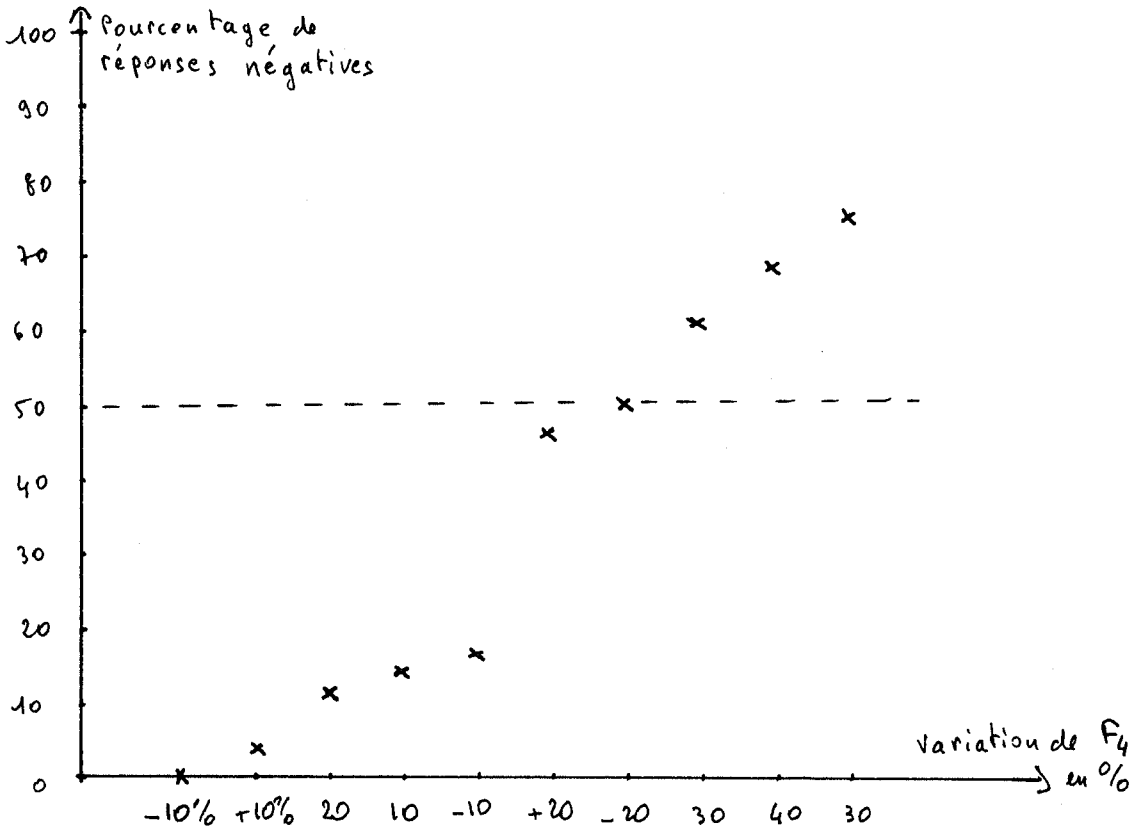


fig 2 : PERTE DE L'IDENDITÉ VOCALE EN FONCTION DE LA VARIATION DE F_4

qu'une faible variation de ce paramètre (de l'ordre de 1%) provoque une différence perceptible. Nous avons cherché à déterminer quelle variation du Fø, toutes choses égales par ailleurs, provoquait la perception que ce n'était plus la même personne qui parlait.

Nous avons réalisé quatre nouvelles synthèses :

- synthèse n° 11 : Fø est augmentée de 10%
- synthèse n° 12 : Fø est diminuée de 10%
- synthèse n° 13 : Fø est augmentée de 20%
- synthèse n° 14 : Fø est diminuée de 20%

Nous avons établi un test qui mêlait, par paires, la voix originale, la synthèse de référence et ces quatre nouvelles synthèses. Ce test a été présenté à des auditeurs avertis, ayant tous une bonne formation de phonéticien, auxquels l'enjeu du test avait été expliqué. Nous avons ensuite eu une discussion avec chacun d'entre eux pour tester le test.

Les résultats sont très cohérents d'un auditeur à l'autre. Le jeu des paires permettait d'avoir une échelle de variation de Fø allant de 0% à $\pm$ 40%, et de fait le nombre de réponses négatives (ce n'est pas la même personne) croît avec l'importance de la variation de Fø. (voir figure 1)

Une rupture nette dans la perception se produit lorsque la variation de Fø atteint 30%. Une variation de 20% provoque des réponses ambiguës, mais restant en-dessous du seuil de 50%. Enfin tant que l'écart est de 10%, il ne provoque pas de perception de variation de timbre.

Il est à noter que la comparaison qui, pour tous les auditeurs ne donne que des réponses positives (c'est la même personne) est celle qui réunit les voix les plus "typées". En effet, l'augmentation de Fø de 10 et 20% donne un timbre plus criard aux synthèses 11 et 13, et donc des voix plus remarquables.

Nous avons d'ailleurs remarqué que quelles que soient les paramètres étudiés jusqu'à présent, une augmentation du paramètre provoquera toujours un effet perceptif plus net qu'une diminution de la même quantité.

Nous sommes conscients des limites de ce test, dont le principal défaut est de n'avoir comporté qu'un seul timbre de référence. Nos auditeurs ont cependant été unanimes à reconnaître qu'il y avait réellement des paires pour lesquelles la réponse négative ne faisait aucun doute. Or la possibilité leur avait été laissée de ne donner que des réponses positives s'ils l'estimaient ainsi, en leur ayant bien expliqué au préalable qu'il s'agissait de modifications faites sur la même voix.

3°) Essais sur le quatrième formant

La même expérience a été reprise en présentant de nouvelles synthèses où la valeur en fréquence du formant 4 variait de 0% à 40%, toutes choses égales par

ailleurs.

Les résultats montrent une progression plutôt linéaire du pourcentage de réponses négatives et ne présentent pas de rupture catégorielle aussi nette que précédemment. Il semble cependant qu'une variation de F4 de 30 ou 40% soit nécessaire pour provoquer la perception de changement d'identité. (voir figure 2)

CONCLUSION

Cette recherche de relations invariantes, caractéristiques d'une voix donnée n'en est encore qu'à ses débuts. Il est évident que c'est par une déformation simultanée et progressive de plusieurs paramètres que l'on obtiendra le passage de la voix d'un locuteur à un autre et non par la variation unilatérale d'un seul paramètre. Et tous les paramètres n'auront pas la même importance dans ce processus. Etablir d'abord des hiérarchies de la contribution au timbre entre différents paramètres, pour des locuteurs donnés (rien ne prouve a priori que ces hiérarchies soient les mêmes d'un locuteur à l'autre), établir par là même les plages quantitatives de variabilité à l'intérieur desquelles il n'y a pas de changement de perception de l'identité du locuteur, tels sont les axes dans lesquels nous allons poursuivre nos travaux.

BIBLIOGRAPHIE

HOLMES (1973)

The influence of global wave form on the naturalness of speech from a parallel formant synthesizer - IEEE, 21

KLATT (1980)

Software for a formant synthesizer - J. Acoust. Soc Am., 67

LIBERMAN, LISKER, DELATTRE, COOPER (1959)

Minimal rules for synthesizing speech, J. Acoust. Soc. Am., 31

LIENARD (1977)

Les processus de la communication parlée - Masson, Paris

ACCES LEXICAL ET RECONNAISSANCE DE GRANDS VOCABULAIRES

J.P. HATON : Equipe Reconnaissance des Formes et Intelligence Artificielle
Centre de Recherche en Informatique de Nancy
B.P. 239 - 54506 VANDOEUVRE les Nancy Cedex

I INTRODUCTION

La reconnaissance mono-locuteur de petits vocabulaires de mots isolés (une centaine de mots) est actuellement un problème assez bien maîtrisé par de nombreux systèmes commercialisés. Le principe consiste en une comparaison globale de formes acoustiques prototypes avec la forme du mot à reconnaître, les variations de durée et de rythme étant composées en général par un algorithme de programmation dynamique [Sakoe, 1971].

Cette approche globale du problème de la reconnaissance de mots ne semble pas pouvoir être étendue à de grands vocabulaires de plusieurs centaines, voire plusieurs milliers de mots. En effet, on est alors confronté à divers problèmes :

- de ressources informatiques nécessaires : taille mémoire pour stocker les formes acoustiques (de l'ordre du millier de bits par mot), puissance de calcul pour effectuer les comparaisons (sauf si l'on dispose d'un processeur spécialisé, éventuellement sous forme de composant VLSI),
- d'apprentissage des formes acoustiques de référence : il est hors de question de faire prononcer par un locuteur tous les mots d'un grand vocabulaire lors d'une phase d'apprentissage. Les systèmes envisagés doivent être multi-locuteurs ou du moins adaptables aisément à un nouveau locuteur. Cela implique une approche au moins en partie phonétique du problème,
- de limitation intrinsèque d'un algorithme de comparaison de formes globales fondé uniquement sur des données acoustiques pour de grands vocabulaires [Keilin, 1981]. Là encore l'utilisation d'indices et de traits phonétiques semble être la meilleure solution.

Les expériences menées au cours des dernières années sur diverses langues ont montré l'intérêt d'une approche phonétique de la reconnaissance de mots.

Le système réalisé pour le finlandais [Kohonen, 1980] permet la reconnaissance mono-locuteur d'un vocabulaire de 1000 mots. Ce système utilise une segmentation et un étiquetage phonétiques, associés à une technique de hash-code pour la recherche des mots. Une version multi-locuteur a été mise en oeuvre sur un vocabulaire de 200 mots [Riittinen, 1981].

Le système réalisé au CRIN [Haton, 1981] pour la reconnaissance mono-locuteur d'un vocabulaire de 200 mots se fonde sur un système de décodage phonétique centi-

seconde utilisant les données fournies par un banc de filtres. Un algorithme de programmation dynamique assure la comparaison du treillis phonétique d'un mot inconnu avec les transcriptions phonétiques stockées dans le lexique. Cette étude préliminaire a servi de base à la réalisation d'un système de reconnaissance de 1000 mots [Mari, 1984] dont nous reparlerons au paragraphe IV. L'extension de ce système à la reconnaissance multi-locuteurs de très grands vocabulaires est actuellement en cours, conjointement avec nos travaux sur un système expert de décodage phonétique [Carbonell, 1984].

Au MIT, Zue et ses collègues ont mené plusieurs expériences intéressantes pour l'anglais [Shipman, 1982], [Zue, 1983] qui concluent également à l'intérêt d'une approche phonétique pour la partition de grands vocabulaires et l'identification de mots.

Nous présentons dans cet article une brève discussion des problèmes posés par la reconnaissance de grands vocabulaires. Le paragraphe II est consacré à l'utilisation de contraintes acoustico-phonétiques pour la description phonétique grossière de mots et la sélection de sous-vocabulaires de petite taille. Le paragraphe III concerne l'accès au lexique et la reconnaissance d'une unité lexicale. A titre d'illustration, nous terminons par une esquisse du système en cours de développement à Nancy.

II DEFINITION DE SOUS-VOCABULAIRES

La prise en compte d'informations phonétiques permet, par une caractérisation grossière en termes de grandes classes phonétiques d'un mot inconnu, de restreindre la reconnaissance à un sous-ensemble très réduit du lexique tout entier. Après cette première sélection, l'identification d'un mot inconnu se fait alors par une comparaison fine avec les mots du sous-ensemble retenu, ce qui allège considérablement les calculs.

Ainsi, pour l'américain, la simple description des mots en suite de consonnes et voyelles permet de définir des cohortes contenant de 1% à 7,5% seulement des mots pour un vocabulaire de 20 000 mots. Une description un peu plus détaillée, avec six classes phonétiques (voyelles, occlusives, nasales, fricatives fortes et faibles et semi-voyelles) conduit à des cohortes dont la taille moyenne est de l'ordre de la dizaine de mots [Shipman, 1982].

Les expériences que nous avons menées pour le français avec un vocabulaire de 1000 mots recoupent ces résultats [Mari, 1984].

On conçoit l'intérêt d'une telle démarche pour augmenter les performances d'un système de reconnaissance de mots. De plus, la détermination des grandes classes phonétiques peut être réalisée à l'aide d'un ensemble de règles d'une façon relativement indépendante du locuteur.

Pour prendre en compte cet aspect multi-locuteur, il est également possible d'utiliser des techniques d'analyse de données pour la sélection de paramètres pertinents. Ainsi Lockwood [Lockwood, 1984] part d'un corpus de 20 locuteurs pour sélectionner quelques paramètres pour les discriminations voisé-non voisé et voyelle-consonne. Il en déduit une description phonétique grossière d'un vocabulaire à l'aide des trois classes : voyelle, fricative sourde et plosive sourde.

III ORGANISATION DU LEXIQUE ET RECHERCHE LEXICALE

Dans un système de reconnaissance de mots à grand vocabulaire le lexique doit être organisé de façon à permettre une recherche rapide des cohortes de mots correspondant au mot prononcé. De façon générale, chaque mot doit pouvoir être accédé en fonction de sa description phonétique grossière et des variantes de cette dernière. La recherche des cohortes peut alors être menée à bien par une technique de hash-code [Zue, 1983], ou bien en définissant a priori pour chaque mot les sous-vocabulaires dans lesquels ce mot est susceptible de se trouver (il y a bien entendu d'importants recouvrements entre les sous-vocabulaires). Une autre technique consiste à organiser le lexique en fichier séquentiel indexé et de calculer une clé d'accès à une cohorte en fonction de la transcription phonétique grossière du mot [Mari, 1984].

Un mot inconnu est ensuite identifié à partir de la cohorte sélectionnée en utilisant des informations phonétiques plus fines. Ceci revient toujours plus ou moins à déterminer une transcription phonétique détaillée de ce mot. Le fait de fonctionner en mode multi-locuteur introduit à ce niveau des difficultés considérables, compte-tenu des variations phonémiques importantes constatées d'un locuteur à l'autre. Une première solution consiste à précompiler un lexique en extension en utilisant un ensemble de règles morpho-phonologiques. Ce lexique est censé contenir toutes les variantes phonologiques du vocabulaire considéré. Cette solution peut être coûteuse en place mémoire dans le cas de grands vocabulaires. Il est également possible de prévoir une représentation des mots du lexique sous une forme arborescente permettant de stocker les variations (insertion, substitution et élision de phonèmes) les plus fréquentes. Dans ce dernier cas, la reconnaissance d'un mot utilise un algorithme de comparaison d'une telle représentation arborescente avec une chaîne ou un treillis dephonèmes. Les techniques dérivées de la programmation dynamique peuvent être utilisées avec profit pour résoudre ce problème [Mari, 1984].

Quoiqu'il en soit, il est nécessaire de progresser dans le domaine du décodage phonétique multi-locuteur pour véritablement être en mesure d'obtenir de bons résultats à ce niveau. Les recherches actuellement en cours permettent d'espérer de nets progrès dans un avenir proche.

IV UN EXEMPLE DE SYSTEME [Mari, 1984]

Pour illustrer notre propos, nous décrivons ci-dessous les principales caractéristiques du système de reconnaissance de 1000 mots fonctionnant actuellement dans notre laboratoire et de ses développements actuels. Ce système utilise comme information acoustique les données fournies par un analyseur spectral à banc de 16 filtres et par un détecteur de fondamental. A partir de ces données brutes sont déterminés, d'une part, les instants de début et de fin de mot et, d'autre part, un ensemble de 22 paramètres qui sont utilisés à la fois pour déterminer la description phonétique grossière d'un mot et pour établir la transcription phonétique détaillée de ce mot.

La figure 1 donne un schéma synoptique du système.

Dans ce système, deux processus sont menés en parallèle :

- la transcription phonétique du mot à reconnaître sous forme d'une chaîne de phonèmes à réponses multiples,
- la description d'un mot en termes de classes phonétiques et la sélection d'une cohorte de mots satisfaisant à cette description.

Ces deux processus interagissent fortement tant pour la segmentation de l'onde sonore que pour la caractérisation acoustico-phonétique du mot. Les décisions prises sont liées aux déductions effectuées par un système expert de décodage phonétique [Carbonell, 1984]. L'intérêt de cette approche réside en particulier dans le fait que les règles d'expertise intégrées dansle système sont, dans une large mesure, indépendantes du locuteur, surtout pour ce qui est de la détermination de grandes classes phonétiques.

La description d'un mot en classes phonétiques se fonde sur un ensemble de six classes :

PS plosive sourde
PV plosive voisée
FS fricative sourde
FV fricative voisée
VY noyau vocalique
VO consonne voisée

En fait, ces six classes ne sont pas systématiquement étiquetées et la philosophie générale consiste à ne retenir que les étiquettes "sûres", au sens d'un facteur de plausibilité. L'utilisation d'un raisonnement par règles permet de confronter de façon itérative les décisions prises pour ne finalement conserver qu'une description phonétique peu affectée par des variations phonologiques. Un même mot peut ainsi résulter en plusieurs descriptions phonétiques grossières, ce qui ne nuit en rien aux performances du système, un mot pouvant se trouver dans plusieurs sous-vocabulaires comme il a déjà été signalé.

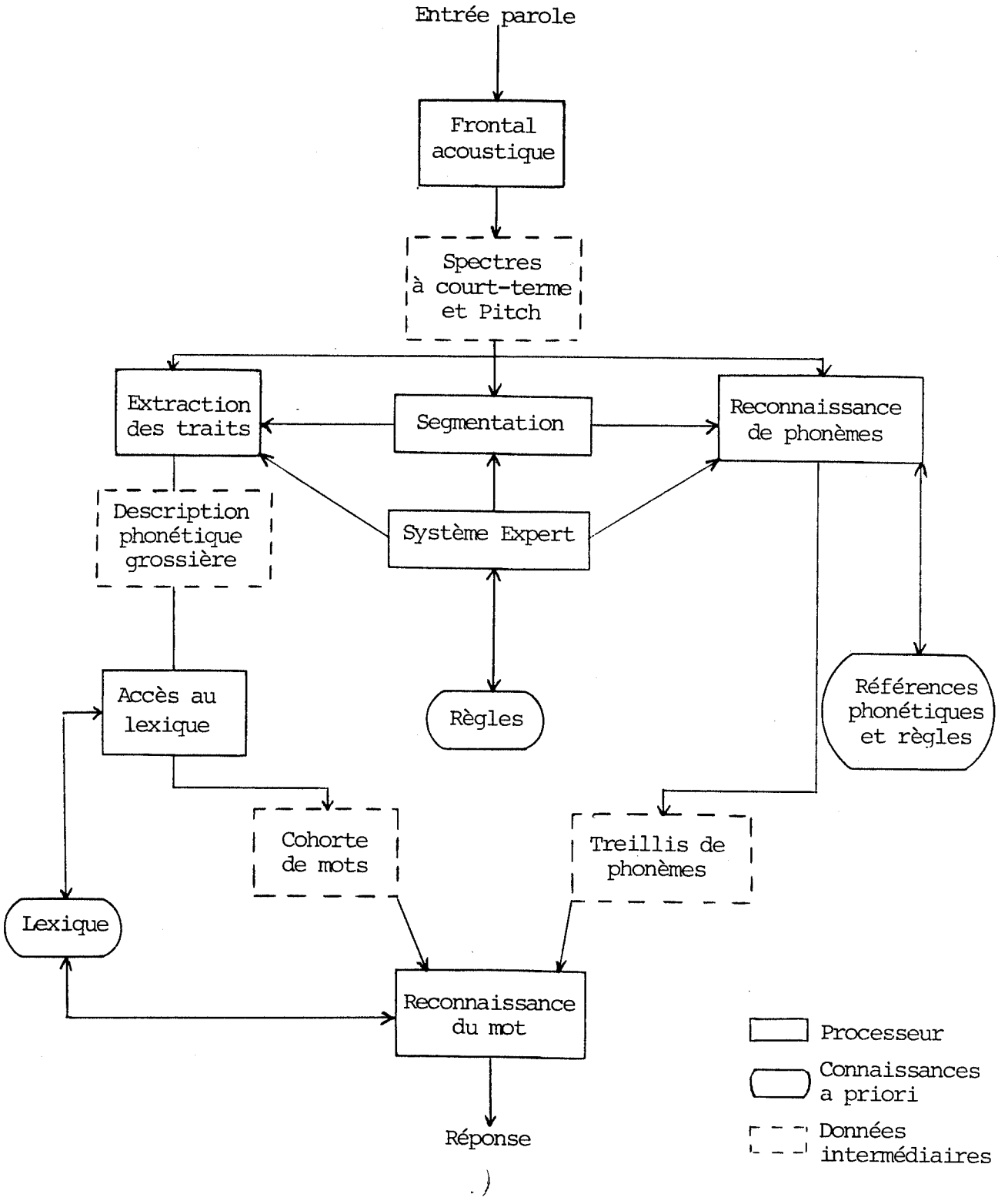


Figure 1 : Schéma général du système

Ainsi, le mot "démontage" /d e m ɔ̃ t a ʒ (ə)/ peut être décrit ainsi :

PV	VY		PS	VY	FV	VY
PV	VY	VY	PS	VY	FV	
PV	VY		PS	VY		
d	e m	ɔ̃	t	a	ʒ	ə

Cette description phonétique sert à sélectionner la cohorte de mots au sein de laquelle une reconnaissance phonétique détaillée est effectuée. La sélection se fait en utilisant la structure de fichier séquentiel indexé du lexique. Une autre organisation est actuellement à l'étude. Avec le vocabulaire de 1000 mots que nous utilisons, environ 400 sous-vocabulaires sont répertoriés, certains ne contenant que 1 ou 2 mots, avec une valeur moyenne de l'ordre de 10.

La reconnaissance phonétique détaillée est ensuite assurée par un algorithme sous-optimal de type programmation dynamique dérivé d'un algorithme que nous utilisons en reconnaissance de parole continue.

Le vocabulaire actuellement utilisé se compose de 983 mots courants, choisis sans tenir compte d'éventuelles similitudes phonétiques. Il est difficile de donner des taux de reconnaissance significatifs sur de tels vocabulaires, néanmoins, les expériences effectuées ont mis en évidence l'intérêt de notre approche, aussi bien pour le temps de réponse que le taux de reconnaissance.

A partir de l'expérience acquise, nous travaillons actuellement à une extension su système de façon à pouvoir accepter de très grands vocabulaires, jusqu'à 10 000 mots environ. Une nouvelle organisation du lexique est à l'étude en vue d'une plus grande efficacité. Par contre, la détermination des descriptions phonétiques de mots sera peu modifiée, compte-tenu de l'approche par règles choisie, en dehors de l'évolution constante du système (modification de règles, ajouts, etc...).

V CONCLUSION

La reconnaissance automatique de très grands vocabulaires (plusieurs milliers de mots) est un problème important au sein du dialogue oral homme-machine, nécessitant la mise en oeuvre de techniques originales.

D'une part, la taille des vocabulaires implique nécessairement que ces systèmes soient multi-locuteurs, ce qui crée des difficultés très grandes dans le décodage phonétique d'un mot.

D'autre part, il est très intéressant de limiter la reconnaissance d'un mot donné à un sous-vocabulaire restreint constitué de mots phonétiquement analogues. Ceci peut être réalisé à l'aide d'une description phonétique grossière des mots en classes phonétiques. Un raisonnement par règles, lié à la notion de systèmes

experts en décodage phonétique, permet d'obtenir une telle description phonétique avec un taux de réussite important.

Nous avons présenté dans cet article, les différents problèmes liés à la mise en oeuvre de systèmes de reconnaissance de grands vocabulaires et nous avons illustré le propos à l'aide du système de reconnaissance de 1000 mots réalisé dans notre laboratoire.

Il reste encore beaucoup à faire dans ce domaine, mais on peut raisonnablement espérer des réalisations importantes dans les années à venir.

REFERENCES

[Carbonell, 1984] Carbonell N., Fohr D., Haton J.P., Lonchamp F. and Pierrel J.M., (1984), "An Expert System for the Automatic Reading of French Spectrograms", Proc. IEEE ICASSP, San Diego.

[Haton, 1981] Haton J.P., (1981), "Reconnaissance d'un vocabulaire de 200 mots utilisant une approche phonétique", Rapport final d'un contrat SEFT.

[Keilin, 1981] Keilin W.J., Rabiner L.R., Rosenberg A.E. and Wilpon J.G., (1981), "Speaker Trained Isolated Word Recognition on a Large Vocabulary", J. Acoust. Soc. Am., Vol. 70, S60.

[Kohonen, 1980] Kohonen T., Riittinen H., Jalanko M., Reuhkala E. and Haltsonen S. (1980), "A Thousand-Word Recognition System Based on the Learning Subspace Method and Redundant Hash Addressing", Proc. 5th ICPR, Miami.

[Lockwood, 1984] Lockwood P., (1984), "Proposition de mots dans un système de reconnaissance de mots isolés multi-locuteurs : une approche en vue du traitement des grands vocabulaires", 13èmes Journées d'Etude sur la Parole, GALF, Bruxelles, 28-30 mai 1984.

[Mari, 1984] Mari J.F. and Haton J.P., (1984), "Some Experiments in Automatic Recognition of a Thousand Word Vocabulary", Proc. IEEE ICASSP, San Diego.

[Riittinen, 1981] Riittinen H., Haltsonen S., Reuhkala E. and Jalanko M., (1981), "Experiments on an Isolated-Word Recognition System for Multiple Speakers", Proc. IEEE ICASSP, Atlanta.

[Sakoe, 1971] Sakoe H. and Chiba S., (1971), "A Dynamic Programming Optimization for Spoken Word Recognition", IEEE Trans. ASSP, n° 26, pp 43-49.

[Shipman, 1982] Shipman D.W. and Zue V.W., (1982), "Properties of Large Lexicons. Implications for Advanced Isolated Word Recognition Systems", Proc. IEEE ICASSP, Paris.

[Zue, 1983] Zue V.W. and Huttenlocher D., (1983), "Computer Recognition of Isolated Words from Large Vocabularies", Trends and Applications IEEE Computer Society, May 1983.

VARIABILITE ET INVARIANCE : L'ESPACE VOCALIQUE EN SWAHILI

J.M. HOMBERT et G. PUECH (Université Lyon 2).

La variabilité interlocuteurs dans le domaine acoustique est un phénomène bien connu (voir par exemple Peterson et Barney (1952) pour l'espace vocalique de l'anglais américain) ; en revanche, la variabilité dans le domaine perceptuel a été étudiée de manière beaucoup moins systématique. Nous nous proposons de montrer ici que des locuteurs possédant, en théorie, le même système phonologique, peuvent opérer un découpage perceptuel significativement différent de l'espace vocalique comme le montre le fait qu'un même stimulus acoustique peut être perçu comme voyelles distinctes par différents locuteurs.

1. Caractéristiques des stimuli synthétiques

Un jeu de 53 voyelles synthétiques a été réalisé en laboratoire. Ces voyelles ont une durée de 250 ms et une mélodie légèrement descendante (de 120 à 110 Hz). L'espace formantique est couvert de la manière suivante : le 1er formant varie entre 250 et 750 Hz avec un pas de 100 Hz ; l'incrément du 2ème formant est de 200 Hz dans un intervalle compris entre 650 et 2350 Hz ; une série arrondie et une série non-arrondie, pour les voyelles d'avant et pour les voyelles d'arrière, ont été synthétisées en faisant varier le 3ème formant (voir fig. 1 et tableau 1).

L'espace vocalique retenu est trapézoïdal , ce qui signifie que les valeurs correspondant aux "coins" inférieurs du carré précédemment défini ont été écartées. De plus, pour ne pas allonger abusivement la durée du test, nous avons éliminé quelques combinaisons formantiques de la zone centrale dans la mesure où ces combinaisons ne paraissaient pas essentielles.

2. Analyse phonologique préliminaire

On procède d'abord à une analyse du système paradigmatique de la langue étudiée. Dans le cas du swahili parlé par les quatre sujets testés (deux hommes et deux femmes parlant le dialecte de Zanzibar), on a, sous l'accent, un système de 5 voyelles, sans opposition de longueur : i, e, a, o, u . On choisit un ensemble de formes qui illustrent les oppositions dégagées en s'assurant

qu'elles conviennent bien pour le dialecte testé. On donne la préférence à des paires minimales de structure syllabique simple si possible. Dans le cas du dialecte swahili de Zanzibar, on a pris les mots suivants :

tita	gerbe
teta	médire
(ma)tata	problèmes
tota	immerger
tuta	butte

Les mots retenus pour illustrer l'ensemble des oppositions du système sont disposés dans un tableau qui comprend en outre une case vide (figure 2.).

3. Tâche des sujets

On familiarise d'abord les sujets avec l'opération qui consiste à isoler une syllabe ou une voyelle dans un mot quelconque. On les entraine alors à isoler la syllabe puis la voyelle pertinente dans les mots choisis pour l'expérience :

Pour [tita] les locuteurs doivent extraire [ti] puis [i],
pour [tata] ils doivent extraire [ta] puis [a], etc.

Les sujets écoutent alors la bande et doivent, après présentation de chaque stimulus, pointer sur la case correspondant à la voyelle dont ils jugent que le stimulus est une réalisation acceptable ; sinon, ils pointent sur la case vide : ils procèdent ainsi à un découpage de l'espace vocalique en fonction de leur organisation perceptuelle.

4. Compte-rendu et analyse de résultats

Les figures 3 à 6 présentent les résultats. Les stimuli identifiés quatre fois sur cinq au moins comme une voyelle déterminée ont été regroupés pour constituer une aire de dispersion perceptuelle. Corrélativement, les stimuli qui, quatre fois sur cinq au moins, ont été jugés comme en dehors du système ont été regroupés dans une aire hachurée.

En observant les figures 3 à 6 on constate :

a) qu'un même stimulus peut appartenir pour un locuteur à l'aire perceptuelle d'une voyelle du système et faire partie de l'aire hors-système pour un autre locuteur.

b) qu'un même stimulus peut appartenir pour un locuteur à l'aire perceptuelle d'une voyelle et à l'aire perceptuelle d'une autre voyelle pour un autre locuteur.

Ainsi, par exemple, les stimuli 31, 24, 26, 19, 32 et 36 sont compris dans l'aire recouverte par |e| pour le sujet S1 et sont jugés hors-système par S3. Le stimulus 30 appartient à l'aire |o| pour le sujet S2, et à l'aire d'exclusion pour le sujet S4. Le stimulus 16 appartient à l'aire de |u| pour S2 et de |o| pour S4. Le stimulus 29 appartient à l'aire de |a| pour S1 et à l'aire de |o| pour S2, S3 et S4.

5. Implications

Il existe un problème de comparaison entre voyelles acoustiquement différentes et perceptuellement identiques. Une première approche consiste à rendre homothétiques les espaces acoustiques de voyelles en définissant des points d'ancrage : points extrêmes de chaque espace considéré ou centre de gravité de ce même espace. On trouvera dans Nearey (1977), Gertsman (1968), Disner (1980, 1983) des études approfondies qui illustrent cette démarche appelée "normalisation". L'approche que nous proposons est différente puisque la normalisation est opérée par les sujets eux-mêmes. Elle consiste en effet à utiliser le filtre perceptuel des locuteurs/allocutaires pour définir des aires pertinentes pour le système, sans faire appel à des paramètres dont on sait qu'ils dépendent en partie de facteurs physiologiques indépendants.

On a ainsi mis en évidence le fait que, pour un même système, des locuteurs pouvaient avoir des aires perceptuelles qui se chevauchent. Le résultat est d'une grande importance pour comprendre comment, en synchronie, un équilibre se maintient entre différents parlers d'une même langue, et comment, en diachronie, les changements phonétiques peuvent s'implémenter. Le fait qu'un même stimulus puisse être assigné à des voyelles différentes par différents locuteurs montre bien que tous les locuteurs n'organisent pas leur perception de la même façon. Dans le modèle développé par Hombert (1984), une des origines possibles des changements phonétiques est le décalage qui peut exister entre le codage de certains locuteurs qui visent une cible X et le décodage opéré par d'autres allocutaires qui interprètent le son perçu comme Y. L'application au swahili de

la méthode exposée ici met en évidence la réalité d'un tel décalage et à ce titre confirme la fécondité de l'hypothèse évoquée précédemment.

Cette expérience suggère par ailleurs qu'il serait souhaitable, pour la synthèse de la parole, de choisir des valeurs formantiques situées en dehors des zones de chevauchement perceptuel.

Références

DISNER S.F. 1980. Evaluation of vowel normalization procedures, Journal of the Acoustical Society of America 67(1), pp.253-261.

DISNER S.F. 1983. Vowel quality : the relation between universal and language specific factors, UCLA Working Papers in Phonetics 58, 158 p.

GERSTMAN L.H. 1968. Classification of self-normalized vowels, IEEE Transactions Audio-Electroacoustics AU-16, pp. 78-80.

HOMBERT J.M. 1979. Universals of vowel systems : the case of centralized vowels, Proceedings of the 9th International Congress of Phonetic Sciences, vol 2, Copenhague, pp. 27-32.

HOMBERT J.M. 1984. Phonétique expérimentale et diachronie : application à la tonogénèse, Th. Doctorat d'Etat, Aix-en-Provence, Université de Provence.

NEAREY 1977. Phonetic feature systems for vowels, Unpublished doctoral dissertation, University of Connecticut.

PETERSON G.E. et BARNEY H.L. 1952. Control methods used in a study of the vowels, Journal of the Acoustical Society of America, 24, pp. 175-184.

PUECH G. 1983. Un fragment de phonologie polylectale, in Principes de grammaire polylectale par A. Berrendonner, M. Le Guern et G. Puech, Lyon, P.U.L. pp. 161-231.

N° des stimuli	F1	F2	F3	N° des stimuli	F1	F2	F3
1	250	2350	3100	28	550	1050	2500
2	250	2150	3100	29	550	850	2500
3	250	1950	2900	30	550	650	2500
4	250	1750	2900	31	650	1950	2900
5	250	1500	2500	32	650	1750	2900
6	250	1250	2500	33	650	1500	2500
7	250	1050	2500	34	650	1050	2500
8	250	850	2300	35	650	850	2500
9	250	650	2300	36	750	1950	2900
10	350	2350	3100	37	750	1750	2500
11	350	2150	3100	38	750	1500	2500
12	350	1950	2900	39	750	1250	2500
13	350	1500	2500	40	750	1050	2500
14	350	1050	2500	41	750	850	2500
15	350	850	2300	42	250	1950	2300
16	350	650	2300	43	350	1950	2300
17	450	2150	3100	44	450	1750	2300
18	450	1950	2900	45	550	1750	2300
19	450	1750	2900	46	650	1500	2300
20	450	1500	2500	47	750	1500	2300
21	450	1050	2500	48	250	850	2700
22	450	850	2500	49	350	850	2700
23	450	650	2500	50	450	850	2700
24	550	2150	3100	51	550	850	2700
25	550	1950	2900	52	650	850	2700
26	550	1750	2900	53	750	850	2700
27	550	1500	2500				

Tableau 1 - Valeurs des formants F1, F2 et F3 pour les 53 stimuli synthétiques (les stimuli 42 à 47 représentent les voyelles antérieures arrondies, les stimuli 48 à 53 représentent les voyelles postérieures non arrondies).

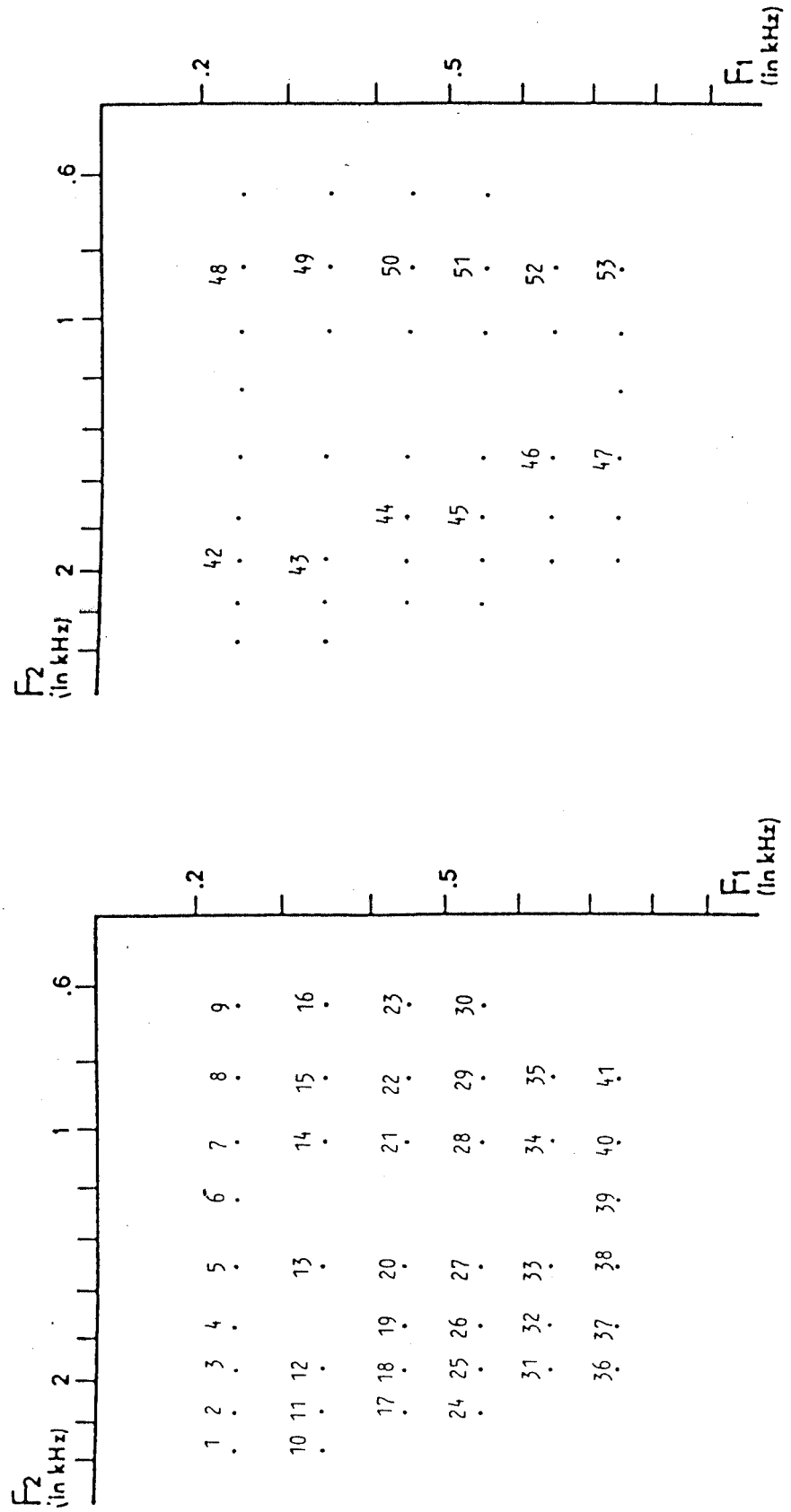


Fig. 1 - Distribution des stimuli synthétiques (voir tableau 1 pour les valeurs de F3 des stimuli 42 à 53).

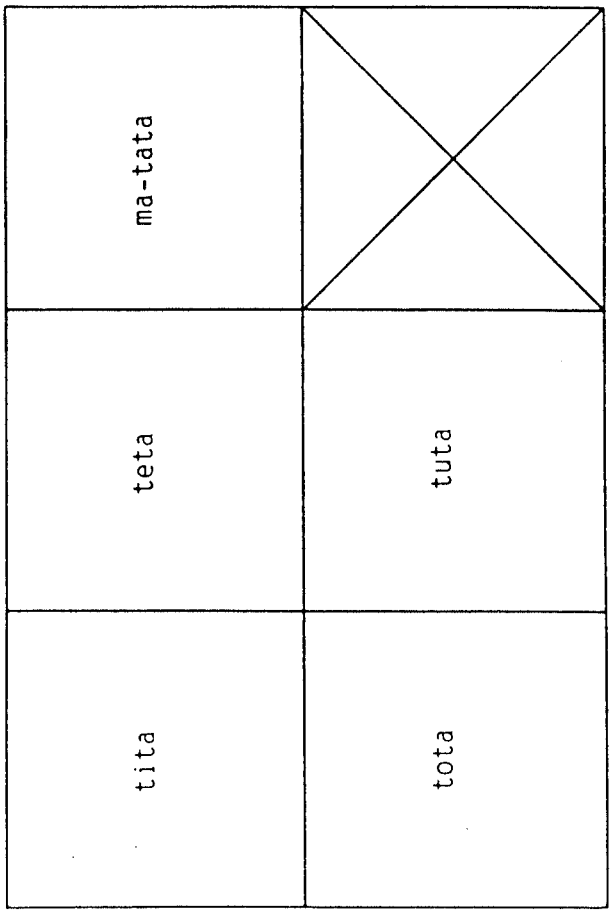


Figure 2 - Choix proposés aux sujets

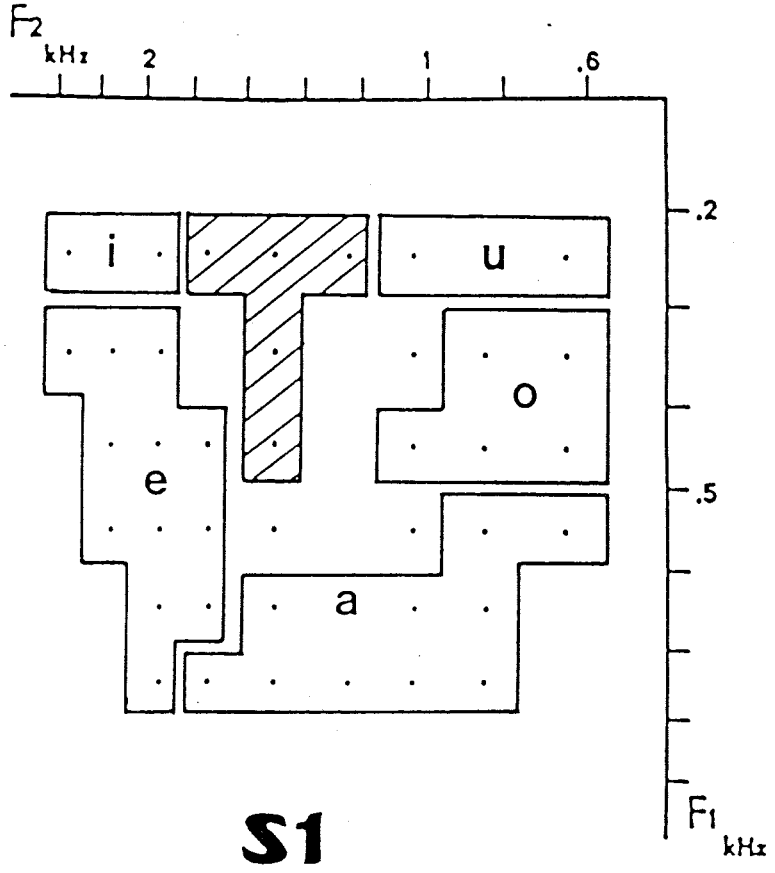


Figure 3 - Découpage perceptuel du locuteur S1
(les stimuli 42 à 53 ne sont pas inclus dans cette représentation)

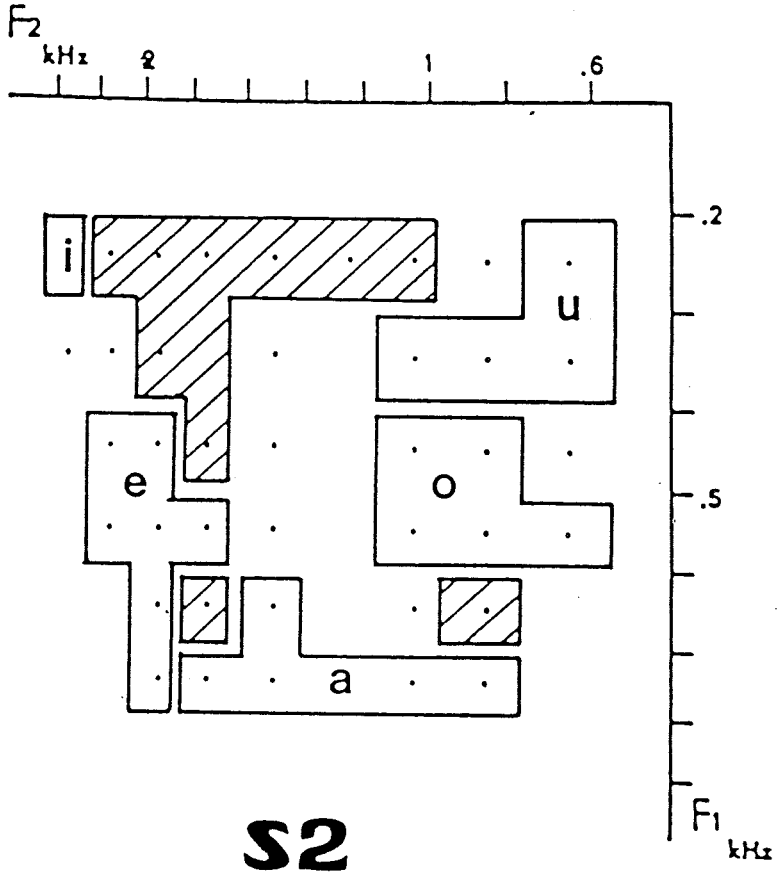


Figure 4 - Découpage perceptuel du locuteur S2
(les stimuli 42 à 53 ne sont pas inclus dans cette représentation)

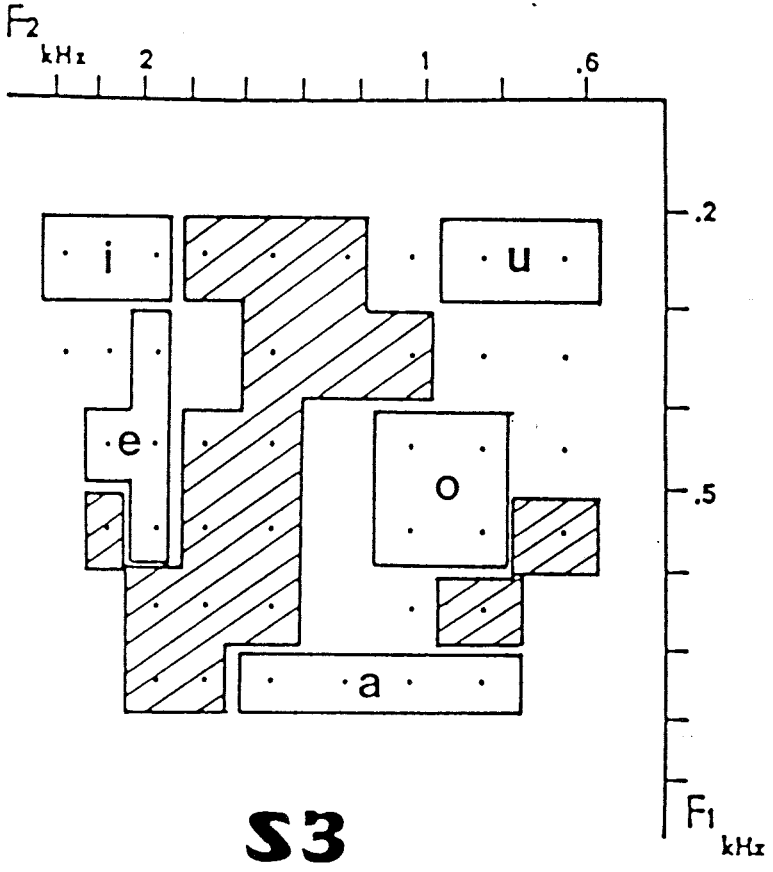


Figure 5 - Découpage perceptuel du locuteur S3
(les stimuli 42 à 53 ne sont pas inclus dans cette représentation)

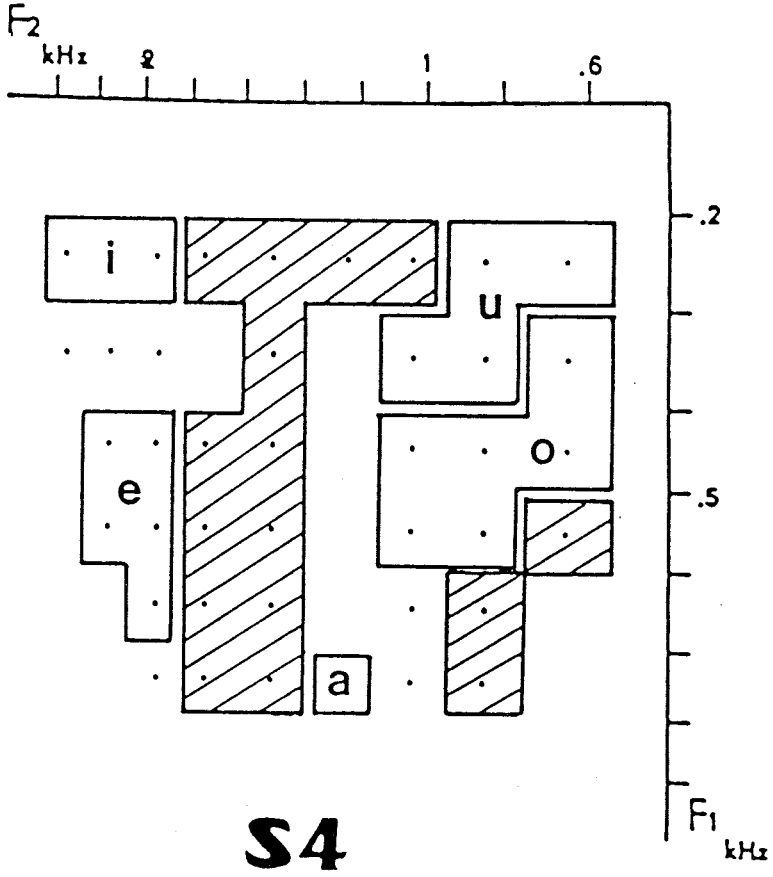


Figure 6 - Découpage perceptuel du locuteur S4
(les stimuli 42 à 53 ne sont pas inclus dans cette représentation)

EXPERIENCES EN RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEUR MULTIREFERENCE

J. MARIANI, M. ESKENAZI, E. MAJANI, A. CARRIER, R. DE FRANCE
LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France

Dans l'optique de réaliser un système capable de reconnaître un vocabulaire indépendamment du locuteur, deux approches ont été envisagées : adaptation au locuteur ou réalisation d'un vocabulaire "multiréférence".

- La première approche (1, 2) consiste à établir un vocabulaire de référence associé à la voix d'un locuteur comprenant les mots du vocabulaire à reconnaître et une phrase-clef choisie pour la variété des phonèmes qui la composent. Pour un nouveau locuteur, l'adaptation se fait par cadrage automatique par programmation dynamique entre la phrase-clef de référence et la réalisation du nouveau locuteur. Il convient alors de déterminer la transformation polynomiale permettant de transformer les références spectrales pour les faire correspondre à la voix du nouveau locuteur. Cette approche nous a conduits à définir des transformations polynomiales différentes suivant la nature des sons analysés (silence occlusif, voyelle, fricative, consonnes voisées), et s'est avérée, de ce fait, délicate.

- La seconde approche a nécessité la prononciation du vocabulaire par une large population (100 personnes), présentant des variétés dialectales et acoustiques importantes. Pour chaque mot, un algorithme de classification automatique a permis de déterminer des types de prononciations qui sont les centroïdes des nuages constitués par des prononciations voisines des mots. La distance utilisée est la distance globale de ressemblance entre les mots, obtenue par programmation dynamique.

Plusieurs références (8 à 12) correspondant à ces centroïdes sont alors conservées. Deux méthodes de classification (UWA - Unsupervised Without Averaging (3), proche des nuées dynamiques (4), et CHA - Classification Hiérarchique Ascendante (5)) ont été utilisées avec des résultats comparables (95% de reconnaissance sur un vocabulaire de chiffres, pour un ensemble de 20 locuteurs n'appartenant pas à la population d'apprentissage).

Ces méthodes sont cependant coûteuses en puissance de calcul. Une approche incrémentale de classification sous-optimale (6, 7, 8) a alors été utilisée. Le vocabulaire est ici prononcé par un premier locuteur. Pour chaque nouveau locuteur, prononçant un mot du vocabulaire, la ressemblance du mot avec les références déjà conservées pour ce mot est calculée. Si cette ressemblance est supérieure à

un seuil, cette nouvelle prononciation est ajoutée à la liste de références pour le mot. Il apparaît que cette méthode augmente sensiblement le nombre de références retenues, mais donne des performances équivalentes. La détermination du seuil est, par contre, cruciale pour éviter la divergence de la méthode.

Une stratégie de correction d'erreur, par reconnaissance de mot-clef et filtrage syntaxique permet une reconnaissance quasi exempte d'erreurs. Le protocole est écrit en VLISP sur PDP 11/23, la reconnaissance étant implantée sur microprocesseur. Des essais utilisant une entrée par canal téléphonique, sont en cours, en conjonction avec un système de synthèse vocale.

(1) M. HUNT
Speaker adaptation for word-based recognition system
ASA Meeting, Ottawa, 1981.

(2) J.P. EMPTAZ
Un système d'adaptation au locuteur pour la reconnaissance
DEA, Ecole Centrale de Paris, 1982.

(3) L.R. RABINER
On creating reference templates for speaker-independant recognition of isolated words
IEEE Trans. on ASSP, vol. 26, n° 1, february 1978.

(4) E. DIDAY, L. LEBART
L'analyse des données
La Recherche, n° 74, janvier 1977.

(5) I.C. LERMAN
Les bases de la classification automatique
Gauthiers-Villars, 1970.

(6) J. MARIANI
Reconnaissance de la parole continue par diphonèmes
Séminaire GALF-GRECO "Décodage Phonétique", Toulouse, septembre 1981.

(7) M. DABOUZ, L. MICLET
Expériences en transmission de la parole à faible débit par vocoder à classification
Séminaire GALF-GRECO "Analyse du Signal", Paris, décembre 1983.

(8) R. CARRE, C. LACOSTE
Constitution d'un dictionnaire de formants à partir de critères perceptifs
Séminaire GALF-GRECO "Analyse du Signal", Paris, décembre 1983.

PROBLEMES ET METHODES EN EVALUATION DE RECONNAISSANCE PHONETIQUE

J.P. TUBACH : E.N.S.T. Dept. SYC - C.N.R.S. ERA 1044. PARIS.

On s'intéresse ici à l'évaluation des performances au niveau phonétique des systèmes de reconnaissance de la parole. Les systèmes qui fournissent, à un moment donné, un résultat en termes de phonèmes représentent un large sous ensemble des systèmes de reconnaissance de parole continue.

Une évaluation fine et pertinente est utile d'abord aux développeurs de tels systèmes, pour évaluer les conséquences de modifications de règles, d'algorithmes ou de paramètres. On peut aussi se proposer de comparer précisément les résultats de plusieurs systèmes sur les mêmes données (dans la base de données du GRECO, par exemple).

Les problèmes posés sont: quels chiffres fournir, et comment les obtenir. Nous pensons que ces chiffres doivent être obtenus automatiquement, sans travail manuel de l'expérimentateur (faute de quoi les évaluations ne seront faites que trop rarement), et qu'ils doivent représenter une quantité d'information modérée, pour être rapidement interprétables. (Mais on ne saurait se contenter d'un "le taux de reconnaissance des phonèmes est de 74,28%", qui est à peu près dénué de signification). Il faut également s'assurer de la validité statistique des résultats, et donc étudier la longueur du corpus de test nécessaire.

Nous proposons de décrire les phrases de test par une transcription phonétique "élargie", permettant de prendre en compte la variabilité permise au(x) locuteur(s) (variantes régionales, pauses, etc). Il importe de se rendre compte que c'est ici que l'on fixe de façon précise ses objectifs au système. Cette information est utilisée pour générer une chaîne de symboles, dite "solution", représentant les différentes solutions possibles, compte tenu d'éléments facultatifs (ex: e muets, pauses).

La méthode est une généralisation de celle proposée dans (1) pour la segmentation. La solution est rapprochée du résultat de la reconnaissance par l'algorithme de Wagner et Fisher (2), qui permet d'appairer les éléments de deux chaînes de façon à minimiser une "distance globale d'édition", qui tient compte du coût fixé pour une insertion, une omission ou une substitution de symbole.

La connaissance des symboles ainsi appairés permet de constituer une matrice de confusion, augmentée d'une colonne pour les symboles "en trop" et de deux lignes pour les symboles "manqués", obligatoires et facultatifs. L'édition de cette matrice peut fournir tous les résultats désirés. Pour une évaluation rapide, ils sont regroupés dans le taleau suivant:

PHONEMES A TROUVER: nn (OBLIGATOIRES: nn FACULTATIFS: nn)
PHONEMES TROUVES : nn

JUSTES OBLIG.	nn	(xx%)	} 100%	JUSTES FACULT.	nn	(xx%)	} 100%
FAUX OBLIG.	nn	(xx%)		FAUX FACULT.	nn	(xx%)	
MANQUES	nn	(xx%)					
EN TROP	nn	(xx%)					

nn et xx représentent des valeurs numériques. Les pourcentages sont rapportés au nombre de phonèmes à trouver, dans chaque catégorie: obligatoire et facultatif.

Pour une évaluation plus détaillée, on peut obtenir une analyse par symbole trouvé (%juste, %faux, et %en trop), et par symbole de la solution (%juste, %faux et, si obligatoire, %manqué). On peut également procéder à des regroupements de phonèmes (ex: voyelles ouvertes et fermées, occlusives etc.)

En ce qui concerne la longueur du corpus, si on cherche à évaluer une probabilité p par des observations donnant une fréquence f dans un corpus de longueur L, la considération de l'intervalle de confiance à 95% conduit à une incertitude relative $er = |p-f|/p = 2\sqrt{(1-p)/p.L}$ Numériquement:

pour $p \sim 70\%$ L=100 --> er=13% L=1000 --> er=4% L=10000 --> er=1,3%
pour $p \sim 5\%$ L=100 --> er=100% L=1000 --> er=30% L=10000 --> er=10%

Les tests sur 1000 phonèmes sont tout juste suffisants. Mais 10000 phonèmes représentent environ 500 phrases, donc un effort très important pour l'acquisition des données et leur traitement.

BIBLIOGRAPHIE

(1) M.D. DI BENEDETTO, J.P. TUBACH. Automatic and computer-aided evaluation of speech segmentation. FASE/DAGA '82, Goettingen, Septembre 82.
(2) R.A. WAGNER, R.J. FISHER. The string to string correction problem. Journal of the ACM vol 21, no 1. Janvier 74.

Remarque: Une partie du travail qui a conduit à cet article a été effectué au Centre Scientifique IBM France, à Paris.

EXPERIENCE DE SEGMENTATION DE PHRASES CONNUES PAR UN SYSTEME EXPERT

J.P. TUBACH : E.N.S.T. Dept SYC - C.N.R.S. ERA 1044. PARIS.
P. DUMOUCHEL : I.N.R.S. Télécommunications. MONTREAL.

Une expérience a été menée à l'ENST, où le système SERAC du CNET (1) a été utilisé pour segmenter la parole continue en unités de taille phonémique, en s'aidant de la transcription phonétique des phrases. Ce type de travail a une application privilégiée en apprentissage automatique.

La démarche adoptée consiste à travailler comme un expert humain examinant le spectrogramme d'une phrase connue: repèrer des segments faciles à identifier (occlusives et fricatives non voisées...), puis poursuivre la segmentation entre ces îlots de confiance en tenant compte du fait que l'on sait ce que l'on cherche (par exemple variation d'énergie dans telle bande de fréquence pour une transition voyelle - consonne nasale).

- Le premier pas consiste à déterminer des segments de 4 grands types : SIL (silence), FRI (fricatives V-), VOC (vocaliques) et IND (indéterminés). Ceci est fait par un automate dont les changements d'état sont commandés par des franchissements de seuils par l'énergie globale et la densité de passages par zéro du signal. (il s'agit d'une version simplifiée d'un algorithme décrit dans (2)).

- A partir de la transcription phonétique de la phrase, on crée sa description en termes de SIL, FRI, VOC, qui est rapprochée des résultats du 1er pas au moyen de l'algorithme de Wagner et Fisher (3), qui permet d'appairer les éléments de deux chaînes en minimisant une "distance globale d'édition".

- On sait alors de quels phonèmes se composent les segments VOC: par exemple /aRa/ dans "caractère" ou au pire /udomyr d / dans "accoude au muret de ce ...". On applique alors des règles spécifiques pour localiser les transitions correspondantes. Nous en avons expérimenté un sous-ensemble, du type:
Voyelle <--> occl. V+ : variation d'énergie entre 500 et 5000 Hz
voyelle <--> cons. nasale: variation d'énergie entre 500 et 2000 Hz
etc...

Les résultats sont sympathiques, mais ne portent pas sur un corpus assez vaste, et des règles supplémentaires restent à ajouter.

Nous estimons cette méthodologie valable, et ce travail sera poursuivi. Mais, sur le plan informatique, l'utilisation d'un gros système expert sur un ordinateur trop peu puissant pour le recevoir pose des problèmes de performances tels qu'ils

annullent les avantages attendus d'un tel système par rapport à la démarche algorithmique classique (facilité de modification de règles et/ou de paramètres pour faire des essais nombreux).

BIBLIOGRAPHIE

(1) D. GILLET et al. SERAC : Un système expert en reconnaissance acoustique - phonétique. 4ème Congrès AFCET-RFIA. Paris, Janvier 84
(2) J.P. TUBACH, M.D. DI BENEDETTO. Two cooperative methods for the segmentation of running speech. FASE- DAGA'82. Goettingen, Septembre 82.
(3) R.A. WAGNER, R.J. FISHER. The string to string correction problem. Journal of the ACM vol 21, no 1. Janvier 74.

UN SYSTEME DE VERIFICATION DU LOCUTEUR

J. MARIANI, J.L. GAUVAIN, J.L. SOURY
LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France.

Les particularités des méthodes de reconnaissance globale, qui effectuent la reconnaissance d'une réalisation acoustique en comparant la forme émise à l'ensemble des formes de référence apprises, font que les paramètres propres à la voix du locuteur sont pris en compte dans le processus de reconnaissance. Cela a pour conséquence de rendre les systèmes dépendants du locuteur, qui doit être le même à l'apprentissage et à la reconnaissance. Dans le cas contraire, les performances de reconnaissance sont dégradées, et ce d'autant plus que les voix sont différentes. Ce défaut peut être utilisé pour réaliser un système de vérification du locuteur, en examinant la note de reconnaissance de la réalisation représentative de la ressemblance entre le mot prononcé et la référence reconnue (1).
Parmi les paramètres utilisables pour reconnaître l'identité vocale d'un locuteur, figurent le timbre, les particularités articulatoires, les particularités phonologiques, la valeur du fondamental et de ses variations, le rythme (2).
Dans les méthodes de reconnaissance globale utilisant des méthodes de programmation dynamique, les méthodes d'analyse usuelles ne donnent pas spécifiquement une analyse du fondamental (utilisée cependant par A. ROSENBERG (4)), et effacent, dans le cadre de l'alignement temporel, les variations rythmiques. A notre sens, cependant, ces paramètres ne sont pas fondamentaux, le fondamental ne permettant de faire qu'une classification grossière entre classes de locuteurs (masculin/féminin par exemple) et le rythme pouvant être modifié par des variations intra-locuteurs. Par contre, les trois premiers paramètres sont présents.

Notre système est entièrement vocal. Lors de l'apprentissage, chaque utilisateur prononce 3 fois son nom, et 3 fois un mot de passe qui lui est propre. Lors de la reconnaissance, le système demande, par synthèse, au sujet de prononcer son nom. Il compare alors cette réalisation aux références qu'il possède pour trouver la plus voisine. La comparaison de la note de reconnaissance à un seuil S1, assez large, entraîne la validation à cette étape, ou une demande de répétition (3 essais sont accordés au locuteur). Lors de la seconde étape, le système demande au sujet de prononcer son mot de passe, et compare sa réalisation aux références correspondant à l'identité déclarée. La comparaison de la note de reconnaissance à un seuil S2, plus sévère que S1, entraîne la validation, ou une demande de répétition.

Cette procédure a l'avantage de permettre une reconnaissance grossière liée à un filtrage tolérant lors de la reconnaissance du nom, donc un processus rapide pour un vocabulaire qui peut être important, puis une reconnaissance fine sur un vocabulaire réduit aux 3 références du mot de passe. Le choix de conserver plusieurs références est dicté par la nécessité de pouvoir adapter dynamiquement des références aux modifications des réalisations (vieillissement de la voix, cas pathologiques). L'utilisation d'un mot de passe propre au locuteur ajoute un niveau de sécurité supplémentaire.

Des tests ont été effectués sur une population de 20 personnes, en supposant que toutes les personnes autorisées étaient également des imposteurs potentiels, connaissant tous les mots de passe des autres locuteurs autorisés, et essayant systématiquement de se faire passer pour eux. Cela a fait apparaître la difficulté de définir les modalités de test, et le choix d'utiliser une méthodologie systématique (comme nous l'avons fait) ou réaliste (3). Les résultats ont fait apparaître un taux d'acceptation d'imposteurs de 2% et un taux de rejet de locuteur autorisé de 3%, en temps réel. Le protocole est écrit en VLISP sur PDP 11/23, la partie reconnaissance étant implantée sur microprocesseur.

Nos travaux actuels portent sur l'affinage de l'analyse acoustique et des indices retenus, et sur les problèmes de détermination des seuils.

(1) R. VIVES
Vérification de l'identité de locuteurs coopératifs, à travers le téléphone, à l'aide d'un système de reconnaissance de la parole
11e JEP GALF, Strasbourg, 1980.

(2) P. CORSI
Reconnaissance automatique du locuteur : présentation générale, méthodologie et expérimentation, perspectives d'application
Thèse D.I., Grenoble, 1979.

(3) G. DODDINGTON
Voice Identification for entry control
Symposium on Voice Interactive Systems (Applications and Payoffs), Dallas, 1980.

(4) A.E. ROSENBERG
Automatic Speaker Verification : a review
Proceedings of the IEEE, vol. 64, n° 4, april 1976.

TRAITEMENTS ET UTILISATIONS DES SPECTRES VOCAUX MOYENS

B. HARMEGNIES : Département de Phonétique et de Psycho-acoustique
Université de l'Etat à Mons

La mise au point d'outils fiables permettant de typer les voix s'avère de plus en plus nécessaire à la résolution de problèmes très diversifiés (1-4). Deux types d'approches président à la recherche d'indices acoustiques caractéristiques de la qualité vocale : l'analyse à court terme (5) (étude de segments sélectionnés) et l'analyse à long terme (recherche de descripteurs de la voix dans sa globalité). Deux orientations constituent ce dernier courant : l'une est caractérisée par le recours aux "contours" de divers paramètres acoustiques (6); l'autre se base sur l'examen des données spectrales à long terme (7). C'est principalement dans cette dernière voie que s'inscrivent nos travaux.

Dans le cadre de notre étude, les moyennes spectrales sont utilisées en tant que descripteurs acoustiques. Celles-ci sont calculées à partir de FFT en 400 points, ce qui confère une grande précision à l'analyse. Les distances entre spectres moyens sont évaluées par plusieurs indices comparatifs dérivés des statistiques descriptives (univariées et multivariées). Ainsi que le montre l'étude actuelle de leur puissance respective, plusieurs de ces indices s'avèrent d'ores et déjà présenter un pouvoir discriminatif comparable à celui des distances classiques. Des études en matière de courbes de langues et d'objectivation de pathologies de la prononciation sont actuellement en cours. Cependant, jusqu'à présent, la procédure a principalement été testée en reconnaissance de locuteurs. L'accent a été mis sur l'étude contrastive des distributions d'échantillonnage des indices mis en oeuvre en situation de comparaison intrasujet et intersujet. Les faibles pourcentages de recouvrement de ces dis-

tributions semblent augurer de résultats prometteurs. Ainsi, lors d'une expérience impliquant 10 locuteurs, les 2 meilleurs indices comparatifs ont fourni, dans 97% à 98% de nos 4.500 comparaisons intersujets, des valeurs qu'aucune comparaison intrasujet n'avait pu produire.

Actuellement mise à l'épreuve dans plusieurs secteurs de la reconnaissance acoustique, la procédure pourrait prochainement être étendue à des domaines aussi divers que l'étude des répercussions vocales de divers types de surdité, l'amélioration des stratégies d'évaluation en pédagogie des langues ou même le diagnostic de certaines pathologies.

(1) PTACEK, P.H., SANDER, E.K., "Age recognition form voice", J.S.H.R., 9, 1966, 273-277.

(2) WEINBERG, B., BENNET, S., "A study of talker sex recognition of esophageal voices", J.S.H.R., 14, 1971, 391-395.

(3) LEVIN, H., LORD, W., "Speech pitch frequency as emotional state indicator", I.E.E.E. Trans., 5, 2, 1975, 259.

(4) CORSI, P., "Speaker recognition : a survey, in HATON, J.-P.,Automatic speech analysis and recognition, Reidel, 1982, 277-308.

(5) SU, L.S., LI,K.P., FU, K.S ., "Identification of speakers by use of nasal coarticulation", J.A.S.A., 56, 6, 1974, 1876-1882.

(6) ATAL, B.S., "Automatic recognition of speakers from their voices", Proc. of the I.E.E.E., 64, 4, 1976, 460-475.

(7) BUNGE, E., "Herkenning van sprekers door een computer", Philips techn.rev., 37, 7, 1977, 179-192.

SHERPA: UN SYSTEME DE RECONNAISSANCE DE LA PAROLE CONTINUE - RESULTATS ET DEVELOPPEMENTS.

A. ANDREEWSKY, F. POIRIER: LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX
M. DESI: CHU - 94 LE KREMLIN-BICETRE
C. FLUHR: Université PARIS-SUD - 91405 ORSAY CEDEX

L'objet de cette communication est de présenter d'une part la méthode de classification à l'apprentissage dans le système SHERPA, d'autre part les expériences de reconnaissance ainsi que les développements envisagés.

1. CLASSIFICATION

A partir de la mise en correspondance des extremums de la courbe moyenne de l'énergie (CME) lissée avec les éléments de la chaîne phonétique, on procède à l'étiquetage d'un sous-ensemble des spectres du corpus.

L'apprentissage s'effectuant sur un gros corpus, on a recours à une méthode de classification en ligne basée sur un algorithme à seuil. La distance spectrale de type convergence uniforme, est calculée sur des spectres identiquement étiquetés.

En fin de classification, on élimine du dictionnaire de références ainsi constitué d'une part les spectres aberrants et d'autre part les spectres non discriminants.

Parallèlement un ensemble d'outils a été développé pour juger la qualité de la classification.

2. EXPERIENCES DE RECONNAISSANCE

De nombreuses expériences de reconnaissance ont été tentées pour tester et modifier la stratégie d'apprentissage.

Les résultats préliminaires ont permis de corriger l'étiquetage et la classification pour constituer un dictionnaire de références plus performant en reconnaissance.

On teste actuellement une segmentation basée sur le lissage de la courbe moyenne de l'énergie et de la courbe de variabilité du signal en déterminant le retard optimal de variabilité automatiquement à l'apprentissage.

Enfin, une meilleure utilisation des informations associées aux références devrait améliorer sensiblement le taux de reconnaissance phonétique.

3. RESULTATS ET DEVELOPPEMENTS

Les efforts porteront désormais sur l'amélioration de l'analyseur phonétique.

En même temps, on s'intéressera à la reconnaissance de mots isolés, sur un gros dictionnaire. Cette expérience permettra de tester des procédures originales d'accès lexical à partir du treillis phonétique.

JONCTION ENTRE UN SYSTEME DE COMPREHENSION DE LA PAROLE CONTINUE ET UN SYSTEME DE RAISONNEMENT DEDUCTIF.

P. FOSSE, P. IVANOFF, J. MARIANI, D. MEMMI
LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France
D. KAYSER
LRI - Bât. 490 - Campus Universitaire - 91405 ORSAY CEDEX - France.

Le système de compréhension de la parole continue ESOPE du LIMSI (1) permet de reconnaître les phrases d'un langage syntaxiquement structuré.

Le système BIDIR, conçu par D. COULON et D. KAYSER (2), permet de répondre à des questions écrites relatives au développement de l'enfant entre 0 et 3 ans, grâce à un raisonnement déductif ayant pour but de faire coïncider par inférences successives de règles l'énoncé (la question) et l'une des connaissances données au système. Ces règles peuvent être strictement descendantes (exemple : si à l'âge x l'enfant commence à faire quelque chose, alors à l'âge y il ne sait pas encore le faire, pour x>y) ou bidirectionnelles (exemple : si l'enfant fait quelque chose c'est qu'il sait le faire, et réciproquement), et chaque règle est affectée d'un coefficient de vraisemblance. Un module complémentaire ("SAUVEUR") ayant pour fonction de permettre l'inférence de règles différant légèrement des règles activables en cas d'échec dans le raisonnement s'ajoute au système BIDIR pour constituer le système "PENSEUR".

Il convenait de remplacer le dialogue à l'aide du clavier par un dialogue entièrement oral. Pour ce faire, le langage d'interrogation a été décrit pour des questions relatives au développement des activités motrices chez l'enfant. La structure du langage (syntaxe et lexique) a été introduite grâce au système ARBUS d'aide à la génération de langage d'application (3). L'utilisateur génère donc de manière interactive, en étant aidé par le système, le langage qu'il souhaite utiliser sous forme de grammaire hors-contexte. Les entités lexicales sont introduites sous leur forme graphémique. Un générateur aléatoire de phrases permet de vérifier la validité du langage construit, et des modifications peuvent facilement être effectuées.

La traduction phonétique des mots introduits sous leur forme orthographique est réalisée automatiquement, avec prise en compte des variantes phonologiques (liaisons, élision du e caduque, ...), à l'aide d'un programme initialement conçu pour la synthèse, et modifié pour s'adapter aux variantes de réalisation phonologiques ou dialectales (4).

Le langage ainsi constitué a un facteur de branchement statique moyen de 5 mots, les phrases ont une longueur moyenne de 8,9 mots, ce qui représente un ensemble de 5010^6 phrases possibles. La perplexité est de 3,1 mots, et la com-

plexité de 12,8 mots. La réponse aux questions reconnues est obtenue par inférences successives de règles, dans le cadre d'une stratégie bidirectionnelle, et tentative d'unification avec les énoncés connus. Elle est affectée d'une probabilité de vraisemblance, qui modifie sa formulation.

Sur un corpus de 20 phrases de test, 15 ont été reconnues au premier essai, et 19 après 2 essais. Les réponses aux phrases reconnues ont toutes été correctes. Le temps de calcul (PL1 sur NAS 9060) est de 10 s pour une phrase de 2 s (4 s pour l'analyse et le décodage phonétique, 1 s pour la reconnaissance, 5 s pour le raisonnement).

Les perspectives sont de lier plus étroitement les stratégies de reconnaissance de la phrase, et de raisonnement, bien que l'approche actuelle (utilisation de la syntaxe pour la reconnaissance, mais pas pour la compréhension) soit en accord avec certaines idées exposées récemment par M. GROSS (5).

(1) J. MARIANI
ESOPE : un système de compréhension de la parole continue
Thèse de Docteur es-Sciences, Orsay, 1982.

(2) D. COULON, D. KAYSER, H. ABIR, J.M. DAVID, J.P. FOURNIER
Modélisation du raisonnement approximatif
Rapport LRI, n° 99, 1980

(3) D. MEMMI, J. MARIANI
ARBUS : a tool for developing Application Grammars
9th COLING, Prague, 1982

(4) F. NEEL, M. ESKENAZI, J. MARIANI
A method to automatically constitute phonetic dictionaries for speech comprehension systems
JASA, vol. 73, sup. 1, 1983.

(5) M. GROSS
La formalisation des langues naturelles
Pour la Science, septembre 1981.

UN COPROCESSEUR DE PROGRAMMATION DYNAMIQUE POUR LA RECONNAISSANCE DE LA PAROLE CONTINUE

G. QUENOT, A. REICHART, J.L. GAUVAIN, J.J. GANGOLF
LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France
J.M. FRAILONG
Ecole Normale Supérieure - 45, rue d'Ulm - 75005 PARIS - France.

Les algorithmes de programmation dynamique utilisés pour la reconnaissance de la parole fournissent d'excellentes solutions aux problèmes d'alignement temporel et de segmentation. Cependant, ces algorithmes nécessitent une quantité de calculs, proportionnelle au carré du nombre d'événements retenus pour représenter les entités à reconnaître, qui ne peut être aisément supportée en temps réel par un microprocesseur standard. Un co-processeur de programmation dynamique capable de supporter la plupart des algorithmes de reconnaissance a été développé au LIMSI.

Cette machine a été spécialement étudiée pour traiter rapidement la partie coûteuse en temps des calculs de comparaison dynamique, c'est-à-dire : les calculs de distances entre deux spectres et les calculs d'équations locales. Sa structure très générale lui permet d'enchaîner automatiquement les calculs d'équation locale pour une référence ou pour une fenêtre dans une référence, de passer automatiquement à la référence suivante et d'effectuer un prétraitement de haut niveau ne transmettant au processeur maître que les résultats de la reconnaissance.

Le processeur maître n'a qu'à charger les références dans son co-processeur, lui envoyer les spectres de la phrase à reconnaître au fur et à mesure qu'ils arrivent et à traiter les résultats que lui renvoie celui-ci.

Le co-processeur gère pour chaque référence une matrice phrase à reconnaître-référence dans laquelle il effectue les calculs de programmation dynamique en progressant d'une colonne à chaque arrivée d'un spectre de la phrase test. Cette solution favorise l'aspect temps réel.

Ce co-processeur comprend une unité de calcul générale, une unité spécialisée pour le calcul de distance, une autre pour les calculs d'adresse, une mémoire de données de 64 kmots de 16 bits, une mémoire de programme de 1kmot de 32 bits, une unité de contrôle qui gère l'ensemble et une interface multibus. Toutes ces unités fonctionnent en parallèle, le temps de cycle est de 200 ns.

L'équation locale et la structure des données sont programmables, il faut de 5 à 6 µs pour effectuer à la fois un calcul de distance avec 16 paramètres et un calcul d'équation locale, ce qui permet de faire de la reconnaissance de

mots enchaînés avec un temps de réponse quasi-nul sur des vocabulaires de 150 à 200 références acoustiques, ce qui correspond à la capacité actuelle de la mémoire.

UN MODELE COGNITIF POUR L'APPRENTISSAGE ET LA RECONNAISSANCE DE LA PAROLE

D. BEROULE : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX

Les recherches effectuées jusqu'à présent en reconnaissance de la parole ont mené à des réalisations admettant l'ordinateur comme outil de base ; l'utilisation de cette machine qui possède un mode de représentation et de traitement de l'information bien spécifique a certainement orienté la façon d'aborder le traitement des formes acoustiques. Alors que des méthodes aussi efficaces que la programmation dynamique fournissent des résultats appréciables, un certain nombre de problèmes fondamentaux demeurent cependant sans solution ; il n'est pas impossible que ces problèmes soient résolus dans l'avenir au moyen de machines présentant une organisation mieux adaptée à l'apprentissage et la reconnaissance de formes, organisation peut-être plus proche du modèle humain.

C'est dans cette optique que nous proposons un modèle fonctionnel qui s'inspire à la fois de données physiologiques (propriétés essentielles des réseaux nerveux) et comportementales (résultats d'expériences de psycholinguistique). Ce modèle repose sur le filtrage et la diffusion d'information à travers un réseau élaboré par apprentissage.

L'apprentissage d'une forme y est fondé sur la stabilisation de certaines structures spécifiques activées automatiquement, et non pas sur l'extraction et le stockage de paramètres en mémoire. La reconnaissance d'une forme se matérialise par la réactivation automatique de la structure qui lui est associée, ce qui n'implique pas l'existence d'une procédure de recherche reposant sur des comparaisons successives de la forme d'entrée avec des formes préalablement stockées. On substitue également à la notion de segmentation celle de filtrage du message acoustique, réalisée au moyen de capteurs spécialisés qui réagissent à certains événements particuliers.

Parmi les caractéristiques du modèle proposé, citons le parallélisme des processus, la hiérarchisation en niveaux d'abstraction qui présentent un comportement similaire, la prise en compte des effets de contexte et de la fréquence des entités, la non-dissociation des processus d'apprentissage et de reconnaissance.

Ce modèle s'accompagne d'une simulation informatique qui permet de tester systématiquement son comportement en fonction de ses paramètres internes (nombre

de niveaux d'activités, facteurs de transmission, seuils d'activation ...) et en fonction des caractéristiques du corpus traité (déterminisme de l'entrée, taille du vocabulaire, complexité de la syntaxe ...). Le programme écrit en langage d'assemblage est constitué de procédures de création de liens entre cellules, de procédures d'activation et de désactivation agissant sur une mémoire qui se structure progressivement en réseau.

EVALUATION DE SYSTEMES DE RECONNAISSANCE GLOBALE ET REALISATION D'UN SYSTEME DE DETERMINATION AUTOMATIQUE DE PERFORMANCES.

J. MARIANI, J.L. GAUVAIN, J.L. SOURY
LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX - France

Tester, ou comparer, des systèmes de reconnaissance de parole nécessite un lourd travail. Il convient de constituer une base de données importante (cf Base de Données "Parole" du GRECO "Communication Parlée" (1)), puis d'effectuer la reconnaissance du contenu de cette base et de décrire les résultats de ces tests en fonction des paramètres que l'on désire étudier. Cela représente plusieurs mois/homme pour chaque système.

Un corpus de parole a été élaboré par le Pannel III/RSG10 de l'OTAN. Ce corpus multilangue est destiné à évaluer les systèmes de reconnaissance de mots isolés ou enchaînés. Il comprend 5200 chiffres prononcés isolément, 2550 suites de 3 chiffres, 1100 suites de 4 chiffres et 1963 suites de 5 chiffres (soit environ 27000 chiffres) prononcés par 19 locuteurs (14 masculins et 5 féminins) dans 4 langues (anglais, français, hollandais, allemand), les locuteurs parlant leur langue naturelle ou une langue apprise (français ou anglais). Le corpus est enregistré sur magnétophone.

Les paramètres étudiés sont : l'effet du sexe (masculin/féminin) et du locuteur, l'influence de la langue et du fait que le locuteur parle sa langue naturelle, l'effet de la longueur de la suite de chiffres, de la position du chiffre dans la suite, de la nature du chiffre, de leur succession et de leur ressemblance, l'influence du prix et de l'encombrement du système testé.

Cinq systèmes de reconnaissance de mots isolés (INTERSTATE-VRM (USA), THRESHOLD-8040 (USA), RSRE (GB), TELEFUNKEN-ADES (RFA), VECSYS-MOISE (France)), et quatre systèmes de reconnaissance de mots enchaînés (VERBEX-1800 (USA), NEC-DP200 (Japon), RSRE (GB), LIMSI-MOZART (France)) ont été testés. Le système VRM a été testé dans deux sites différents (USA et GB). Des tests comparatifs ont également été faits avec des auditeurs humains (à TNO, Hollande).

Afin de pouvoir comparer les systèmes correctement, il était demandé d'ajuster les seuils de rejet aussi haut que possible, et de ne pas retoucher les paramètres du système après la phase d'apprentissage.

Les tests effectués pour les mots isolés montrent des résultats corrects (mieux que 95%) pour la plupart des systèmes, et une relative indépendance aux paramètres étudiés, pourvu que l'apprentissage soit suffisant (plusieurs références par mot, ou moyenne). Les tests effectués pour les mots enchaînés présen-

tent par contre pour certains systèmes, et en particulier lorsque l'apprentissage a été fait uniquement en mot isolé, des résultats parfois surprenants (jusqu'à 100% d'erreurs sur les suites de 5 chiffres pour certains locuteurs !).

Les résultats obtenus par MOISE sont de 99,1% sur les chiffres isolés. Ceux de MOZART sont de 98,8% sur les chiffres isolés, de 94,5% sur les suites de 3 chiffres, 92,6% sur les suites de 4 chiffres et de 85,4% sur les suites de 5 chiffres (soit 94,5% de moyenne sur les entités à reconnaître, et 97,4% sur les chiffres isolés ou enchaînés).

Pour les tests en mode isolé, un programme de dépouillement automatique a été réalisé, qui met en correspondance automatiquement, par programmation dynamique, la liste des chiffres prononcés et la liste des chiffres reconnus. Cela permet de tenir compte des reconnaissances excédentaires (bruits parasites, hésitations, toux ...) et des reconnaissances déficitaires (mot non reconnu, trop court, trop long, prononcé trop faiblement ou trop fort). Cet algorithme permet donc de déterminer automatiquement le nombre d'erreurs, et leur emplacement. La généralisation de cet algorithme à la reconnaissance de mots enchaînés est à l'étude.

(1) R. CARRE, R. DESCOUT, M. ESKENAZI, J. MARIANI, M. ROSSI
The French Language Database : Defining, Planning and Recording a Large Database
IEEE ICASSP - San Diego, March 1984.

(2) J.M. BAKER, D.S. PALLETT, J.S. BRIDLE
Speech Recognition Performance Assessments and Available Databases
IEEE ICASSP - Boston, 1983.

SYSTEME DE RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEURS POUR UN VOCABULAIRE DE 130 MOTS. INTEGRATION DANS UN POSTE DE TRAVAIL.

B. FLOCON, Ph. LOCKWOOD, J. SAP

Laboratoires de Marcoussis - CR de la C.G.E.
Route de Nozay - 91460 MARCOUSSIS - France.

INTRODUCTION

L'état actuel des travaux en reconnaissance de la parole permet d'envisager l'intégration d'un système de reconnaissance de mots isolés dans un poste de travail. En effet, les performances sont suffisantes pour pouvoir apporter à l'utilisateur l'avantage de pouvoir utiliser la voix pour entrer certaines commandes.

LE SYSTEME ET SON INTEGRATION

La taille du vocabulaire que nous avons choisi est de 130 mots. Une application du type poste de travail peut tout à fait être déterminée avec un tel nombre de mots ; au delà, le nombre de commandes vocales est trop important pour que l'utilisateur puisse les avoir présentes à l'esprit à un instant donné.

Le type d'entrée vocale que nous avons retenu est celui des mots isolés. La principale raison est qu'un tel système pourra rapidement être intégré dans le poste de travail. Les mots enchaînés ou mieux encore le discours continu sont le moyen le plus naturel de dialogue, mais le vocabulaire utilisable reste de taille limitée (quelques dizaines de mots [1]).

De récents travaux [2] ont montré que l'on pouvait obtenir des performances comparables en reconnaissance de mots isolés, que le système soit mono ou multilocuteur, ceci pour un vocabulaire de 100 - 200 mots.

Les utilisateurs considèrent souvent la phase d'apprentissage comme une opétion contraignante ; il semble donc intéressant de pouvoir disposer d'un système de reconnaissance indépendant du locuteur. La principale critique fait à ces systèmes est que le vocabulaire est figé ; dans le système que nous proposons, l'utilisateur aura la possibilité d'étendre le vocabulaire en associant au système multilocuteur une extension monolocuteur.

* Etude financée en partie par un contrat de la C.C.E. (Programme ESPRIT).

Le système de reconnaissance que nous avons implanté est celui décrit dans [3]. La partie analyse comporte un filtrage à 3400 Hz, un échantillonnage à 8000 Hz et le calcul toute les 16 millisecondes de 9 coefficients cepstraux et d'un terme représentant l'énergie du signal.

Le système dispose d'un ensemble de références obtenu après analyse statistique d'un grand nombre de représentants de chaque mot du vocabulaire provenant d'un échantillonnage varié de locuteurs (50 hommes et 50 femmes).

L'algorithme de reconnaissance calcule une mesure de similarité entre le mot inconnu et tous les éléments de l'ensemble de référence à l'aide de la programmation dynamique.

Un corpus de travail a été réalisé afin de pouvoir effectuer l'analyse statistique avec suffisamment de confiance. Un programme interactif permet d'enregistrer tous les mots du vocabulaire avec contrôle immédiat de leur qualité : vérification de la bonne détection début-fin de mots (correction possible s'il y a présence d'un bruit de lèvres par exemple), possibilité d'écoute, de réenregistrement, de visualisation du signal temporel et d'affichage du spectre.

Le signal temporel est stocké simultanément sur bande analogique et sur disque après numérisation. La modularité des programmes permettra de tester le système avec différents paramètres d'analyse, différents algorithmes de classification...

CONCLUSION

Nous disposons à l'heure actuelle d'un système de reconnaissance de mots isolés multilocuteur dont l'intégration sur un poste de travail est prévue assez rapidement. Le vocabulaire sur lequel nous travaillons est formé des commandes d'une machine de traitement de textes. Une extension monolucuteur est prévue afin d'étendre le vocabulaire à la demande de l'utilisateur.

REFERENCES

[1] C. GAGNOULET M. COUVRAT
Seraphine : A connected word speech recognition system
ICASSP 1982

[2] J.G. WILPON - L.R. RABINER - A. BERGH
Speaker independent isolated word recognition using a 129 word airline vocabulary
J.A.S.A. 72 (2) August 1982

[3] B. FLOCON - N. BRIANT - Syril -
4ème congrès AFCET Reconnaissance des formes - Paris - Janvier 1984

UN CORPUS CONTINU POUR L'ETUDE DES VOYELLES DU FRANCAIS

M. ESKENAZI : LIMSI-CNRS - B.P. 30 - 91406 ORSAY CEDEX.

Dans le cadre des recherches sur la reconnaissance automatique de la parole, un des buts à long terme est de pouvoir traiter une parole naturelle. Afin d'arriver à ce but, il est nécessaire de travailler sur un corpus qui s'approche autant que possible de la réalité. Ce n'est cependant pas une tâche facile. Pour effectuer des études acoustico-phonétiques rigoureuses, il est nécessaire de contrôler le contexte des sons recherchés aux niveaux phonétique et prosodique, et les conditions acoustiques de la prise de son. Il faut également que toutes les combinaisons de contextes recherchées aient été prononcées par les locuteurs. Pour ces raisons, la parole spontanée ne peut que rarement être utilisée. La prononciation de ces énoncés doit alors être provoquée artificiellement.

Jusqu'alors les moyens utilisés pour les études sur les voyelles ont fourni des éléments assez éloignés de la réalité :

voyelles isolées - sans la prise en compte de la coarticulation ;

voyelles dans une phrase porteuse - sans une structure prosodique naturelle. Pour tenter de s'approcher d'une parole continue et naturelle, il est possible de présenter un court texte, à lire trois fois - chaque fois sous une forme légèrement différente aux locuteurs. Ces passages fourniront des séries de triplets de différence minimale (du point de vue phonétique, par exemple : jardin, sapin, matin).

La création d'un texte implique de fortes contraintes sémantiques au moment de la recherche lexicale pour qu'un même contexte puisse accueillir trois énoncés différents avec assez de "sens" chaque fois pour que le locuteur n'ait pas d'hésitations au cours de la lecture. Une forte contrainte existe également au niveau syntaxique où les mots d'un triplet doivent tous appartenir à la même classe (verbe, nom, etc) pour pouvoir se présenter à la même place à l'intérieur d'un groupe rythmique. Un dictionnaire de rime est un bon outil pour établir des listes de mots fournissant les contextes désirés. La lecture du texte par plusieurs locuteurs fournit une première base pour évaluer sa validité. Sa transcription en Alphabet Phonétique International en donne une seconde pour l'examen des contextes. La lecture à haute voix, toute en étant une forme non spontanée du discours, est un pas en avant par rapport aux méthodes traditionnelles. Elle permet la constitution de corpus de parole naturelle pour l'étude d'éléments où différents

aspects du contexte doivent être contrôlés.

LA DESCRIPTION PARAMETRIQUE DES PATHOLOGIES LARYNGEES AU NIVEAU ACOUSTIQUE

J. SCHOENTGEN : Institut de Phonétique - Université Libre de Bruxelles.

Le but de l'étude que nous présentons est double:

1) Décrire de manière objective, quantitative et reproductible, sur base d'un ensemble d'indices, les manifestations des pathologies laryngées au niveau acoustique.
2) Développer une procédure automatisée pour extraire ces mêmes indices à partir du signal de parole émis par le locuteur.

Notre travail est subdivisé en deux parties. La première est consacrée à l'étude des voyelles soutenues, la seconde à l'étude de la parole continue (sous forme de phrases isolées). Dans cet article nous résumons les résultats obtenus pour les voyelles soutenues. L'extraction des indices de dysphonie s'opère en quatre étapes:

1) acquisition et numérisation du signal de parole, 2) prétraitement,
3) extraction de l'indice acoustique, 4) évaluation du pouvoir de discrimination de l'indice.

L'étude complète porte sur approximativement 140 locuteurs dysphoniques et un nombre équivalent de locuteurs normaux. Tous les sujets ont été enregistrés à l'aide d'un même équipement (microphone, magnétophone à modulation AM) et dans un environnement acoustique semblable (chambre audiométrique). Les signaux de parole ont été numérisés à 10 kHz (12 bits) et stockés sur des fichiers ordinateur. Les signaux provenant de locuteurs masculins et féminins ont été analysés séparément. Le rôle du prétraitement consiste à transformer le signal d'acoustique, en vue, soit de lui enlever la contribution des cavités supra-glottiques, soit d'amplifier les aspects du signal propres au larynx. Dans ce qui suit nous désignerons un signal issu d'un tel prétraitement comme "auxiliaire". Il est sous-entendu qu'une application préalable d'une procédure de prétraitement augmente le pouvoir discriminant des indices calculés, comparé à l'onde de parole non traitée. Nous avons pris comme point de départ de nos propres recherches les publications de St. B. Davis (1976). Celui-ci fait appel au signal d'énergie résiduelle comme signal auxiliaire. Une évaluation critique des performances de discrimination des attributs

extraits, montre que les indices d'apériodicité réalisent des performances raisonnables sur les signaux provenant de locuteurs masculins par opposition aux résultats médiocres obtenus sur les voix féminines. L'adoption du "critère de variation d'amplitude" (Jospa et Schoentgen, 1982; Jospa, 1984) nous a permis d'obtenir une séparation entre sujets féminins malades et normaux d'une qualité bien supérieure à celle réalisable à partir du signal d'énergie résiduelle; les performances de discrimination augmentent également légèrement pour les locuteurs masculins.

Une autre question à laquelle nous avons tenté de donner une réponse expérimentale est celle du choix de la voyelle la plus susceptible de répercuter, dans le signal de parole, des anomalies au niveau du larynx. D'une étude comparative portant sur les voyelles orales /i/, /e/, /a/, il ressort que la voyelle /a/ l'emporte. Le pouvoir de discrimination s'avère être fonction de l'intervalle en fréquence qui sépare la fondamentale du premier formant.

Quant aux types d'indices acoustiques nécessaires à la caractérisation de la voix dysphonique, l'expérience nous suggérait une classification selon trois grands axes: les indices d'apériodicité en amplitude et fréquence fondamentale, le rapport signal sur bruit, ainsi que le rendement vocal (Schoentgen, 1984). Ce choix a trouvé ultérieurement sa justification au sein d'un modèle phonétique proposé par J. Laver (1981), qui définit le registre modal comme celui pendant lequel le mouvement des vraies cordes vocales est périodique, efficace et sans bruit de friction audible.

Pour évaluer quantitativement les performances de discrimination des indices calculés nous avons eu recours à des méthodes d'analyse typologique. A l'aide de celles-ci nous tentons de séparer, sur base des valeurs des indices extraits, les locuteurs normaux des locuteurs dysphoniques. Si cette tentative se montre fructueuse, l'indice en question est considéré comme discriminatif. Cette approche nous semble supérieure aux méthodes qualitatives qui font appel à des comparaisons avec les niveaux physiologiques ou perceptifs.

Jospa P., Actes des 13e JEP, GALF, Bruxelles, 28-30 mai 1984.
Davis St. SCRL Monograph 13, Santa Barbara, California, 1976.
Laver J., Cambridge University Press, Cambridge, 1980.
Schoentgen J., Speech Comm., 1, 1982.
Schoentgen, J., J. of Speech and Hearing Research (à paraître)
Jospa P. et Schoentgen J., Actes Réunion FASE/DAGA, 1982

ORGANISATION D'UNE BASE DE CONNAISSANCE ACOUSTIQUE ET PHONETIQUE

N. VIGOUROUX et J. CAELEN

Laboratoire CERFIA, Université Paul SABATIER, 118 Route de Narbonne, 31062 TOULOUSE Cedex - France.

Les conclusions des récents travaux en reconnaissance automatique de la parole continue multilocuteur montrent que les problèmes essentiels se situent plus dans l'organisation, dans l'interprétation des données acoustiques et phonétiques, et dans l'extraction des connaissances que dans les méthodes d'analyse et de reconnaissance.

Dans ce cadre, nous avons décidé d'élaborer une Base de Connaisssance (BC) construite automatiquememt à partir d'informations acoustiques [1] et statistiques. Les données acoustiques sont prélevées à trois niveaux d'analyse: échantillon spectral, unités phonétiques, phones (segments courts et homogènes issus de la segmentation automatique [1], [2]).

La BC comporte une base de faits et une base de règles. La base des faits contient la suite de tous les phones réduits:

- par définition de macro-traits pseudo-phonétiques selon la répartition des phonèmes dans l'espace des paramètres (indices spectraux, contextuels, ...),
- par obtention de "pattern matching" par regroupement de phones, en classes au sens d'une métrique donnée.

Quant à la base des règles, elle contient:

- les règles synthétiques de description des unités obtenues par des opérations booléennes et par détermination d'une distance de ressemblance,
- les règles de "savoir" issues des analyses statistiques et de l'étude de l'inférence des règles descriptives.

A chaque nouvelle analyse de corpus, nous distinguons deux traitements, la mise à jour (ou la création lors de l'initialisation du système) et la réduction (exploitation) de ces deux bases.

La BC permet donc de répondre aux objectifs posés dans [1]:

- d'une part, décrire les unités de son choix (phonème, syllabe, mot, ...) pour divers jeux de paramètres,
- d'autre part, traduire des oppositions entre unités phonétiques ou pour une même unité des oppositions entre contextes ou locuteurs.

Nos premiers essais portent sur l'analyse d'un corpus de 27 logatomes trisyllabiques prononcés par cinq locuteurs. Les seules informations prises en compte sont les indices spectraux [3], et ce pour l'unité phonème.

Dans la limite de notre expérimentation, l'analyse de la BC montre l'existence d'invariants acoustiques autant au niveau des zones stables des phonèmes que des phases transitoires entre ces derniers [2].

Au vu de ces résultats, il semble donc possible d'obtenir des règles stables pour la reconnaissance phonétique ainsi que pour la synthèse.

[1] - J. CAELEN, N.VIGOUROUX, G.PERENNOU
Structuration des informations acoustiques dans le projet A.R.I.A.L.
SPEECH COMMUNICATION - Vol 2, No 2-3, July 1983.

[2] - N. VIGOUROUX
Décodage acoustique-phonétique de la parole continue multilocuteur:
Elaboration d'une base de connaissance.
Thèse de 3ème cycle Informatique - TOULOUSE Janvier 1984.

[3] - J. CAELEN, G. CAELEN-HAUMONT
Indices et propriétés dans le projet ARIAL II.
Actes séminaires "Encodages et Décodages phonétiques".
GALF-CNRS - TOULOUSE 1981.

RELATION ENTRE LE TIMBRE DE LA VOIX ET L'ONDE GLOTTIQUE OBTENUE PAR FILTRAGE INVERSE

P.J. PRICE : CNET TSS/RCP - 22301 LANNION

Le timbre de la voix peut marquer des différences individuelles, sociales, linguistiques, géographiques ou de sexe. Il sert aussi à distinguer les voix "normales" des voix pathologiques ou synthétiques. L'étude a pour but de mieux comprendre ces différences pour pouvoir synthétiser des voix plus naturelles et plus variées, en particulier des voix féminines.

La méthode de filtrage inverse employée ici consiste à faire d'abord une approximation des fréquences des formants par DFT. Les zéros sont introduits de manière à supprimer tous les formants sauf un, successivement pour F1, F2 et F3. Les formants sont ainsi quasiment isolés ce qui permet de mesurer avec précision leur fréquence dans le domaine temporel pendant l'excitation principale de chaque période glottique. Ces mesures sont utilisés dans le filtrage final où tous les formants sont supprimés par des zéros corréspondants. Enfin, on tient compte des effets de rayonnement aux lèvres en supposant qu'il correspond à une intégration.

Cette méthode a été employée pour analyser deux réalisations des syllabes [ɛ], [ø], [ʌ], [pʌ], [bʌ] prononçés par huit locuteurs (4 hommes, 4 femmes). Les résultats ont été examinés dans les domaines temporel et fréquentiel. On a trouvé des différences statistiquement significatives entre les hommes et les femmes dans la forme de l'onde glottique et dans son spectre. En particulier, on a constaté une grosse différence liée au sexe du locuteur dans l'onde glottique aux alentours de la fermature des cordes vocales. Ce résultat sera discuté en termes perceptifs et physiologiques, en se référant au modèle à deux masses d'Ishizaka et Flanagan.

STANDARDISATION DE SIGNAUX TESTS POUR L'EVALUATION DES ANALYSEURS DE MELODIE

Ph. MARTIN, Experimental Phonetics Laboratory, New College
University of Toronto, 300 Huron Street,
TORONTO, Ontario, Canada. M5S 2X6

L'évaluation des performances d'un analyseur de mélodie est complexe, et se traduit généralement par le degré de fiabilité et de précision de la mesure de la fréquence fondamentale (pourcentages d'erreurs grossières et fines) par rapport à une courbe de référence). Les erreurs grossières se caractérisent par une mauvaise détection du fondamental (sous-, sur- ou non-harmonique; irrégularités de source); les erreurs fines portent sur la mesure de la fréquence fondamentale supposée correctement détectée.

L'évaluation d'un analyseur X se fait habituellement par comparaison, selon les pourcentages d'erreurs grossières et fines, avec un instrument de référence R opérant sur le même signal de parole S. L'analyseur R peut être de type physiologique (mesure des vibrations laryngées), acoustique (le cepstre constitue une référence très utilisée), perceptif (le jugement est fait sur de la parole resynthétisée à partir des courbes de F_o obtenues par X et R). On conçoit que les performances comparatives d'un analyseur donné X dépendent essentiellement du choix du signal S, dont les caractéristiques particulières pourront mettre en valeur les qualités - ou les défauts - de l'appareil testé ou du système de référence.

Pour réaliser des évaluations plus objectives, qui ne dépendent ni d'un analyseur de référence ni d'un corpus - test aux caractéristiques vagues et difficilement reproductibles, on propose un système de standardisation du signal d'analyse, basé sur la génération de classes de signaux synthétiques reproduisant des caractéristiques spécifiques et bien définies du signal de parole.

Exemples de caractéristiques testées:

Source (mode de vibration laryngée):	- creak
	- modal
	- falsetto
	- souffle
	- etc.
Conduit vocal (bruit intrinsèque):	- occlusives
	- fricatives
	- transitions V-NV/NV-V
	- etc.
Bruit extrinsèque:	- canal de transmission
	- etc.

Une telle standardisation permet la constitution de classes de signaux aisément reproductibles, permettant une évaluation des performances des analyseurs sur des caractéristiques contrôlables du signal.

MODELISATION NON-STATIONNAIRE DE SEGMENTS DE PAROLE : APPLICATION A LA COMPOSITION DE MESSAGES

M.C. CHEVALIER, G. CHOLLET, Y. GRENIER

Département Systèmes et Communications, ERA 1044
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault - 75634 PARIS Cedex 13

Le spectre du signal de parole présente des résonances: le choix d'une modélisation AR(Auto-Régressive) ou ARMA(Auto-Régressive à Moyenne Ajustée) se justifie puisque de tels modèles possèdent des pôles. Cependant, une des principales caractéristiques de ce signal est sa non-stationnarité. Le moyen souvent utilisé pour tenir compte de cette propriété consiste à considérer le signal stationnaire sous une fenêtre de courte durée (vingt à trente millisecondes) et, par conséquent, à estimer un modèle ne dépendant pas du temps. La fenêtre est ensuite décalée par pas de dix millisecondes. Mais une telle méthode présente des inconvénients non négligeables: les modèles estimés peuvent être redondants entre eux si l'analyse s'effectue dans une zône stable du signal; inversement, si la fenêtre contient des évènements transitoires (dans le cas des sons plosifs par exemple), l'estimation des modèles est biaisée.

Une autre approche permet de remédier à ces inconvénients; les variations temporelles du signal sont explicitement introduites dans le modèle qui est basé sur l'hypothèse suivante: une décomposition linéaire sur une famille de fonctions dépendant du temps et connues à priori représente les prédicteurs autorégressifs ou les coefficients de la partie MA /1/. Le problème de la modélisation d'un signal scalaire non-stationnaire est alors ramené à celui de la modélisation d'un signal vectoriel mais stationnaire; il peut être facilement résolu à l'aide d'algorithmes du type Levinson; ceux-ci permettent de déterminer les poids de la combinaison linéaire représentant chaque coefficient à temps variable.

L'utilisation de tels modèles pour la synthèse de la parole nécessite le choix d'une base de fonctions; nous en avons retenu trois: polynômes de Legendre, séries de Fourier, fonctions sphéroïdales aplaties. L'ordre du modèle ainsi que le nombre de fonctions dans la base peuvent être estimés à l'aide de critères tels que celui d'AKAIKE /2/.

Le signal de parole est découpé aux minima de l'énergie en segments, chacun de ces segments étant représenté par un modéle AR non-stationnaire. Il

est à noter que le modèle AR est souvent suffisant (l'information est portée par les formants qui se traduisent par des résonances dans le spectre), et est d'un emploi plus facile que celui du modèle ARMA (la linéarité du modèle AR entraîne la simplicité et la rapidité des algorithmes d'estimation).

Lors de la restitution, nous synthétisons chaque segment, les différents segments obtenus étant ensuite concaténés entre eux.

Les premières expériences paraissent montrer que la qualité des phrases synthétisées est correcte. Les résultats obtenus au moyen de cette technique sont semblables à ceux résultant de l'estimation d'un modèle AR stationnaire, fenêtre par fenêtre; mais les modèles non-stationnaires offrent l'avantage supplémentaire de pouvoir être codés avec moins de bits.

D'autre part, la concaténation des segments, se produisant dans les parties les moins audibles du signal, ne semble pas poser de problèmes majeurs.

REFERENCES.

/1/ Y.Grenier: Time-Dependent ARMA modeling of non-stationnarity signals. IEEE Trans. ASSP, vol.31, n°4, pp 899-911, 1983.

/2/ H.Akaïke: A new look at the statical model identification. IEEE Trans. Aut. Contr. , vol.19, n°6, pp 716-723, 1974.

CLASSIFICATION DES SONS AU MOYEN DE LA PREDICTION LINEAIRE ET D'UN MODELE DU SYSTEME AUDITIF PERIPHERIQUE

Y. DOLOGLOU, J.M. DOLMAZON : Institut de la Communication Parlée
ENSERG, 23, avenue des Martyrs
38031 GRENOBLE Cedex - France.

La présente communication concerne une méthode de reconnaissance de parole qui utilise deux systèmes différents de prétraitement: un modèle du Système Auditif Périphérique (S.A.P.) [1], et un étage à Prédiction Linéaire.

L' intérêt principal de la méthode que nous avons développée se situe dans l'originalité de la méthode de reconnaissance des phonèmes, ainsi qu'en l'utilisation d'un modèle S.A.P. au niveau du codage de la parole.

Le modèle S.A.P. que nous avons utilisé, est composé de deux étages principaux: le premier, qui est linéaire, modélise les mouvements de la membrane basilaire, le deuxième, qui est non linéaire, modélise le comportement des récepteurs sensitifs et des fibres nerveuses associées. Le modèle possede 32 canaux ce qui permet de travailler avec des signaux dont le spectre s'etend de 25 Hz à 4500 Hz.

Ce modèle a été testé dans différentes conditions (régime harmonique, impulsionnel et trapézoïdal, stimulation à deux tons...etc). Le comportement observé confirme que cette version simplifiée du modèle complet, développé au Laboratoire, simule avec une bonne approximation la plupart des phenomènes importants observés dans les fibres du nerf auditif (non linéarité, selectivité, suppression à deux tons ...etc).

Dans un premier temps, nous avons considéré 29 classes de phonèmes. Chaque phonème a été codé en utilisant soit le modèle S.A.P. avec extraction des paramètres physiquement significatifs à partir des réponses, soit la prédiction linéaire. Chaque phonème, mis sous forme d'un vecteur, a été rapporté dans un espace de dimensions appropriées. On a donc créé un support statistique de représentants phonémiques, constitué par 29 nuages (classes) de points. Les nuages ainsi créés se récouvrent. Pour mieux les séparer on a appliqué l'Analyse Discriminante, les classes étant considérées deux par deux. Les 29 classes permettent donc de définir 406 axes discriminants associés avec les projections des centres de gravité de chaque nuage.

La procédure de reconnaissance consiste à projeter le vecteur du phonème inconnu sur les 406 axes discriminants définis pendant l'apprentissage.

° (Supporté par la Fondation Alexander S. Onassis)

Sur chaque axe on calcule les distances entre la projection du point inconnu et celles des centres de gravité. Ces distances sont ensuite accumulées et à la fin de la procédure, la classe qui présente la plus petite distance accumulée est considérée comme correspondant au phonème inconnu.

La méthode proposée a été testée en fonction du nombre d'exemplaires phonétiques par classe ainsi qu'en fonction du nombre de paramètres utilisées pour le codage de chaque phonème. On a constaté que la reconnaissance s'améliore quand le nombre d'exemplaires par classe augmente. Par ailleurs le rôle du nombre de paramètres n'est pas indépendant du nombre d'exemplaires phonétiques par classes: Il apparaît qu'un corpus mal défini (peu d'exemplaires) favorise un petit nombre de parametres, tandis qu' avec un corpus bien défini (beaucoup d'exemplaires) on peut utiliser un plus grand nombre de paramètres. Ces conclusions sont applicables aussi bien pour le modèle S.A.P. que pour la prédiction linéaire.

Nous avons ensuite testé de la parole bruitée (S/B=0 db). Les résultats sont nettement meilleurs en utilisant le modèle S.A.P. qu'utilisant la prédiction linéaire.

Enfin un taux de reconnaissance a été établi à partir d'une phrase de parole continue, (durée voisine de 10 sec), prononcée par un locuteur inconnu. Le modèle S.A.P. a donné des résultats comparables à ceux obtenus par prédiction linéaire. Nous avons obtenu des scores voisins de 75%.

REFERENCE

[1] J.M. DOLMAZON, M. BOULOGNE (1982)
Interaction phenomena in a model of mecanical to neural transduction in the ear.
Speech Communication, 1, 55-73

UTILISATION D'UN MODELE AUDITIF POUR LA RECONNAISSANCE DES OCCLUSIVES

B. DELGUTTE : Centre National d'Etude des Télécommunications
LANNION - France

Pour appliquer les études sur l'audition à la reconnaissance automatique de la parole, certains auteurs ont remplacé le module d'analyse acoustique d'un système de reconnaissance par un modèle d'oreille sans changer les stades ultérieurs du système. Les performances obtenues avec ces modèles n'étaient pas toujours meilleures que celles des méthodes d'analyse traditionnelles. On vise à améliorer ces performances en développant des méthodes de reconnaissance spécifiquement adaptées à la représentation du signal fournie par les modèles d'oreille, simulant ainsi l'adaptation réciproque entre le cerveau et l'oreille. Nous commencerons par localiser les évènements acoustiques importants pour la reconnaissance de la parole en exploitant le fait que les réponses des fibres du nerf auditif privilégient certains phénomènes transitoires.

Cette approche est illustrée par la reconnaissance du lieu d'articulation des occlusives dans des syllabes CV prononcées par des locuteurs masculins et féminins. Après avoir traité ces syllabes par un modèle du système auditif périphérique, on détecte la détente du bruit d'explosion et le début de la partie vocalique par une méthode décrite antérieurement. La reconnaissance est basée sur les représentations spectrales fournies par le modèle en échantillonnant le signal immédiatement après ces deux évènements. On définit un petit nombre de traits acoustiques caractérisant la forme du "spectre auditif" à la détente, et le changement spectral entre la détente et le début de la voyelle. On compare trois méthodes de reconnaissance. La première utilise ces traits d'une façon totalement indépendante du contexte, tandis que la seconde les utilise en supposant le contexte vocalique connu. La troisième méthode exploite le fait que les valeurs de certains traits sont fortement corrélées avec certaines caractéristiques de la voyelle. On peut ainsi adapter les seuils de classification au contexte sans identifier explicitement la voyelle. Des résultats préliminaires pour les occlusives vélaires suggèrent.

que la première méthode donne des taux d'erreur au moins deux fois plus élevés que la seconde, mais que la troisième méthode n'est pas beaucoup plus mauvaise que la seconde. Ces résultats sont en faveur d'une notion d'invariance généralisée qui ne résiderait plus dans un seul indice acoustique, mais dans une procédure d'intégration de plusieurs indices répartis dans le temps et dépendant du contexte.

L'INFORMATION ACOUSTIQUE AU NIVEAU DU NERF AUDITIF

M. BOULOGNE : Institut de la Communication Parlée. L.A. 368 - I.N.P.G. - ENSERG - 23, rue des Martyrs - 38031 GRENOBLE Cedex.

Nous analysons à l'aide d'un modéle non linéaire du système auditif périphérique (DOLMAZON, BOULOGNE, 1982), l'information acoustique présente sur les différentes fibres du nerf auditif.

L'information acoustique du nerf auditif est analysée dans les domaines fréquentiel et temporel. Nous avons étudié la réponse fournie en régime stationnaire à des signaux de parole de synthèse :

- pseudo voyelle à deux formants
- sons non voisés de type [f], [s], [ch]

Dans le cas des voyelles stationnaires, l'information est codée d'une part au niveau de l'activité moyenne des fibres simulées. L'information spectrale est conservée. D'autre part, les paramètres du signal original se retrouvent en étudiant uniquement les taux de synchronisation des fibres. De plus la fréquence fondamentale est présente sur les fibres de fréquence caractéristique élevée.

Dans le second cas (sons fricatifs), l'information est codée de manière très différente. Il n'existe pas de phénomène de synchronisation des fibres. L'information se retrouve uniquement codée par l'activité moyenne des fibres.

La modification des paramètres du modèle nous permet de montrer que les informations contenues dans l'activité moyenne des fibres peuvent disparaître et de ne rester apparentes qu'en analysant leurs taux de synchronisation . Nous montrons également qu'il est possible de privilégier certaines caractéristiques du signal étudié par un choix judicieux des paramètres de calcul du

modèle. Nous montrons qu'en fonction du niveau d'entrée du signal, l'activité moyenne des fibres simulées ne permet plus de retrouver d'information organisée et véritablement significative. Par contre les taux de synchronisation ont conservé leur information.

Nous avons ajouté à notre modèle, un module d'adaptation tel que celui décrit par SMITH et ZWISLOCKI (1975) afin de pouvoir étudier des sons non stationnaires comme les consonnes. Ce module permet en effet de reproduire le régime transitoire du système auditif périphérique. Nous avons utilisé comme signal d'excitation de notre modèle des consonnes de synthèse dans un environnement de voyelles de synthèse. Deux types de consonnes sont étudiées : les consonnes voisées telles que [b], et les consonnes non voisées comme [p].

Il est à noter que tous nos résultats précedents sont en bonne conformité avec les résultats physiologiques observés chez l'animal.

J.M. DOLMAZON, M. BOULOGNE "Interaction phenomena in a model of mechanical to neural transduction in the ear",Speech Communication I (1982) 55-73.

R.L. SMITH, J.S. ZWISLOCKI "Short term adaptation and incremental responses of single auditory nerve fibers", Biol. Cybernetics 17 (1975) 169-182.

UTILISATION DES SPECTRES A LONG TERME POUR DEGAGER DES PROPRIETES ACOUSTIQUES DES LANGUES: ETUDE COMPARATIVE ET DEVELOPPEMENTALE

P. HALLE, B. de BOYSSON BARDIES, L. SAGARD
LIMSI du CNRS (Orsay) et Laboratoire de Psychologie de la MSH.

Dans la production de la parole, les positionnements phonatoires et/ou articulatoires se traduisent dans le domaine fréquentiel. Les spectres à long terme reflètent de nombreux facteurs tels que la qualité vocale propre au locuteur (Bordone-Sacerdote 1969), le matériel verbalisé, l'effort vocal, etc...

Ils traduisent également des caractères spécifiques aux langues (types de phonation, répertoire phonétique). Ce dernier point cependant n'a été que peu étudié (Fant 1973, Tarnoczy 1956, Esling 1983) et parfois controversé.

Nos travaux portent sur l'étude des spectres à long terme de locuteurs masculins et féminins français, algériens et cantonnais, ainsi que sur des spectres à long terme d'enfants de 8 et 10 mois de même appartenance linguistique.

Nous cherchons à déterminer :

a) S'il est possible de dégager à partir des spectres à long terme d'adultes des tendances spécifiques aux langues et de les séparer des facteurs propres au locuteur.

b) Si de telles tendances se retrouvent déjà dans les spectres à long terme des enfants de 8 et 10 mois. Cela appuyerait l'hypothèse d'une mise en place précoce de certains caractères acoustiques propres à la langue-cible dés le stade du babillage.

Les résultats obtenus semblent valider l'intérêt de cette méthode pour des études comparatives interlangues chez des adultes comme pour des études du développement de la parole.

2 types de spectres à long terme ont fait l'objet de notre étude :

- des spectres intégrant l'énergie aux différentes fréquences sur la totalité d'une production vocale
- des spectres ne conservant que les parties nettement voisées.

Un lissage de type cepstral a ensuite été effectué pour éliminer les accidents non pertinents. Les statistiques utilisées pour obtenir spectres à long terme moyens et écart-types sont basées sur une distance rms entre spectres.

SYNTHESE A DIPHONES: VERS UNE MEILLEURE QUALITE ...

J.P. LEFEVRE, E. VIARA : Laboratoire Traitement de Parole, CIT-Alcatel
10bis, rue Louis Lormand - 78320 LA VERRIERE

Afin de permettre la synthèse de vocabulaires non limités, les techniques basées sur la concaténation d'éléments acoustiques minimaux, ont été l'objet de nombreux efforts au cours des dernières années. De telles techniques supposent une analyse préalable permettant la constitution d'une base d'unités. Plusieurs solutions (mot, syllabe, demi-syllabe, élément correspondant à la réalisation d'un phonème, diphone) ont été explorées. Pour de multiples raisons l'approche diphone est apparue séduisante et a fait l'objet de travaux importants en particulier pour le français (1). Malheureusement, à l'heure actuelle, la qualité obtenue demeure encore globalement insuffisante sauf pour certaines applications très ponctuelles. Certes, l'évaluation de la qualité en synthèse est un problème fort délicat. Toutefois l'appréciation de cette qualité depend essentiellement de deux composantes qui sont l'intelligibilité (notion de nature relativement objective) et l'agrément d'écoute (notion plus subjective).

Nos travaux dans le domaine de synthèse par concaténation de diphones ont pour origine les recherches menées par le CNET (1). Dès le début de notre étude, les résultats obtenus nous ont permis d'évaluer la parole synthétisée comme étant de faible intelligibilité, tout en demeurant relativement agréable. Toutefois parler d'agrément, et plus encore essayer de l'améliorer, pour un signal peu intelligible, est relativement illusoire. Nos efforts ont donc porté tout d'abord sur l'amélioration de l'intelligibilité. Bien que les deux composantes de la qualité ne soient pas indépendantes, il nous a semblé commode de lier l'intelligibilité essentiellement au dictionnaire de diphones, et d'associer agrément et prosodie.

Afin de générer un dictionnaire de diphones de bonne qualité, nous avons été amené à proposer un certain nombre de solutions aux problèmes rencontrés lors des différentes étapes (2). Rappellons simplement ici les trois points sans doute les plus importants de notre analyse. Tout d'abord, le corpus à utiliser pour l'extraction des diphones doit être défini avec soin. Ensuite, le choix du locuteur et plus encore son élocution sont particulièrement critiques. Des tests effectués sur de nombreux locuteurs ont mis clairement en évidence combien la qualité de la synthèse était dépendante de leur façon de prononcer. A ce niveau, la mise en place d'une méthodologie nous est rapidement apparue indispensable; en effet tout défaut introduit lors de l'acquisition du corpus n'est

pratiquement jamais récupérable dans la suite de la chaine de synthèse. Finalement la cohérence du dictionnaire doit être assurée. Elle ne saurait l'être si les différentes phases de la réalisation s'étendent sur une trop longue période. L'utilisation d'un outil de génération semi-automatique accélérant le processus est donc absolument nécessaire.

La mise en oeuvre des idées exposées ci-dessus nous a conduit à des résultats très encourageants. Le premier dictionnaire que nous avons généré à partir d'une voix masculine, nous permet de restituer une parole synthétique parfaitement intelligible. De nombreux points de détails encore perfectibles ayant été observés durant la réalisation de ce premier essai, nous envisageons la génération d'un nouveau dictionnaire dans un avenir proche. Ce travail n'est plus une tâche interminable, puisque la méthodologie et l'outil de génération semi-automatique que nous avons développés, nous autorisent la réalisation aisée de dictionnaires dans des délais très courts (de l'ordre d'un mois). Signalons enfin que cet outil nous permet d'apporter une solution jugée pour l'instant satisfaisante au problème lié à l'aspect essentiellement monolocuteur de ce type de synthèse.

A partir d'une synthèse désormais intelligible, nos efforts vont maintenant porter essentiellement sur l'agrément, que nous comptons améliorer en retouchant les règles et patrons prosodiques actuellement implantés dans notre système de synthèse. Dans un premier temps, il est envisagé de vérifier systématiquement, avec l'aide d'un important corpus, les hypothèses proposées par F. Emerard (1). Le développement d'un outil informatique adapté à cet objectif devrait également nous permettre d'associer aisément à chaque dictionnaire les schémas prosodiques correspondant exactement au locuteur ayant prononcé le corpus d'où sont extraits les diphones. En effet, actuellement, l'utilisation de schémas de référence déformés par l'utilisation d'une règle de trois pratiquée sur le fréquence laryngienne moyenne ne nous donne pas entière satisfaction. Par la suite, et suivant les résultats de cette première phase, la mise au point de règles plus complexes, en particulier pour le traitement du rythme, est envisagée...

(1) Courbon J.L., Emerard F., "SPARTE : A text-to-speech machine using synthesis by diphones", IEEE-ICASSP, Paris, Mai 1982, pp.1597-1600

(2) Chollet G., Galliano J.F., Lefèvre J.P., Viara E., "On the Generation and Use of a Segment Dictionnary for Speech Coding, Synthesis and Recognition", IEEE-ICASSP, Boston, Avril 1983, pp. 1328-1331.

MODIFICATION DES PARAMETRES PROSODIQUES EN ANALYSE-SYNTHESE MULTI-IMPULSIONNELLE

M. STELLA: Centre National d'Etudes des Télécommunications
22301 LANNION - France.

La technique d'analyse multi-impulsionnelle (B.S. Atal, Int. Conf. ASSP, pp 611-614, Paris, 1982) s'est révélée très efficace pour obtenir de la parole synthétique d'excellente qualité dans des applications de type transmission (analyse-synthèse). Cette technique, dérivée de la prédiction linéaire, s'affranchit, par une méthode d'optimisation, de la détection du voisement et du calcul de la période fondamentale ; elle permet donc d'éviter les nombreux inconvénients liés aux difficultés du suivi du fondamental. Cependant, cet avantage est diminué par le fait que les paramètres d'excitation obtenus ne sont plus indépendants des paramètres du filtre de prédiction, ce qui fait que la technique multi-impulsionnelle est mal adaptée à des applications de type synthèse de la parole à partir du texte, pour lesquelles on veut pouvoir manipuler les paramètres prosodiques, notamment la mélodie.

On propose ici deux méthodes de modification des paramètres prosodiques (durée et FO) sur des phrases analysées par la technique multi-impulsionnelle. Ces méthodes partent d'une analyse synchrone des phrases, c'est-à-dire qu'on a marqué pour chaque phrase, à la main, le début de chaque période fondamentale. La fenêtre d'analyse multi-impulsionnelle recouvre donc exactement une période dans les régions voisées, et a une valeur constante dans les régions non voisées. Des tests informels ont montré que la qualité de la parole obtenue en analyse-synthèse par cette méthode est meilleure que si on utilise une fenêtre constante.

La première méthode de modification des paramètres prosodiques part de la constatation que, dans les zônes stables tout au moins, on peut répéter les trames d'analyse (coefficients LPC et impulsions) sans détériorer la qualité de la parole de synthèse de manière appréciable : on peut ainsi allonger les parties voisées, en gardant le pitch constant. On peut aussi modifier simultanément la "vitesse"

d'élocution et la mélodie en jouant sur la longueur de la fenêtre de synthèse par rapport à celle de la fenêtre d'analyse, et en comprimant ou en étendant linéairement la suite des impulsions. En combinant ces deux opérations, on peut changer le rythme et le pitch indépendamment. Cette méthode est cependant très lourde, même si elle donne de bons résultats.

La deuxième méthode vise à être beaucoup plus souple, au prix d'une baisse de qualité. On remarque que l'arrangement des impulsions tend à devenir périodique sur les zônes stables des parties voisées. Pour chaque zône stable, on choisit alors la trame dont l'arrangement est le plus représentatif de cette périodicité, et on fait subir à cet arrangement une permutation circulaire, de manière à amener l'impulsion la plus intense en tête. On répète alors cet arrangement sur toutes les trames voisines, en ajustant le gain en fonction de l'amplitude de la plus grande impulsion de chaque trame, et en ajustant sa longueur à la longueur de la trame.

Chaque trame est alors pourvue d'un schéma d'excitation propre. La synthèse s'effectue de manière asynchrone, en utilisant ce schéma d'excitation à la place des impulsions isolées que l'on utilise habituellement.

Les résultats obtenus par cette méthode sont moins bons que ceux de la première méthode, et quelques problèmes subsistent, notamment pour le contrôle de l'énergie de la parole synthétisée. Néanmoins, cette méthode semble prometteuse pour améliorer la qualité de systèmes de synthèse par diphones en prédiction linéaire.

LA SYNTHESE DE L'ARABE A PARTIR DU TEXTE

A. MOURADI, A. RAJOUANI, M. NAJIM
L.E.E.S.A. - Faculté des Sciences - B.P. 1014 RABAT - Maroc.

La langue arabe est caracterisée par un nombre de voyelles restreint(6) et la prédominance des consonnes (29). Comparativement a d'autres langues, l'arabe se distingue par les consonnes d'arrière (pharyngales, emphatiques, uvulaires).

Cet article est relatif à la réalisation d'un système de syntèse de l'arabe à partir du texte au moyen d'un dictionnaire de diphones.
Un tel système recquiert deux composantes essentielles :

- transcription graphème-phoneme.
- constitution d'un dictionnaire de diphones et introduction des faits prosodiques.

TRANSCRIPTION GRAPHEME-PHONEME

La transcription est accomplie par des procédures d'étude de caractéres, équivalentes aux régles de prononciation. Chaque procédure transforme, aprés étude de contexte, un caractère ou une chaîne de phonèmes. Les chiffres sont transcrits d'une façon similaire. En arabe, il y a équivalence entre les sons consonnatiques et les lettres de l'alphabet (exception faite pour les lettres hamzées)surtout au niveau du traitement des voyelles et des effets de liaison, et au niveau de la transcription des chiffres.

La saisie du texte orthographique à traiter se fait à partir d'un terminal intelligent bilingue (arabe-latin). Le texte balayé au moyen d'une fenêtre glissante de largeur variable est analysé graphème par graphème. La transcription est accomplie dès que la largeur de la fenêtre permet l'application d'une procédure d'étude de caractéres. Le programme de transcription écrit en Fortran IV permet de traiter tout texte arabe completement voyellé et comprenant eventuellement des chiffres avec un taux d'erreur presque égal à zéro.

ELABORATION DU DICTIONNAIRE DE DIPHONES

Dans la présente version les diphones sont extraits à partir de logatomes placés dans une phrase porteuse

Chaque diphone est mémorisé sous forme d'un ensemble de paramètres (coefficients de reflexion, gain, indice de voisement) obtenus à partir d'une analyse par prédiction linéaire. Des résultats préliminaires satisfaisants sont obtenus avec un nombre limité de diphones.

L'étude des faits prosodiques (rythme, intonation, accent) est encore à un stade préliminaire.

VERS UNE SYNTHESE PAR LA METHODE DES POLES ET ZEROS

G. FENG: Institut de Phonétique de Grenoble - Institut de la Communication Parlée
LA CNRS 368.

I - INTRODUCTION

La fonction de transfert du conduit vocal peut être décrite par une structure pôle/zéro (P/Z) directement exploitable dans un modèle de synthèse. Nous présentons ici un outil permettant d'étudier les conséquences perceptives dues à la modification d'une répartition P/Z donnée.

Ce synthétiseur P/Z allie souplesse et commodité d'utilisation; il met en jeu un signal d'excitation qui tient compte de l'interaction source/conduit vocal.

II - STRUCTURE DU SYNTHETISEUR

L'éditeur de pôles et zéros utilise une tablette graphique permettant l'acquisition de leurs positions dans le plan-Z. Celles-ci sont stockées en mémoire et revisualisées; au besoin, elles peuvent être aisément modifiées grâce à un jeu de curseurs.

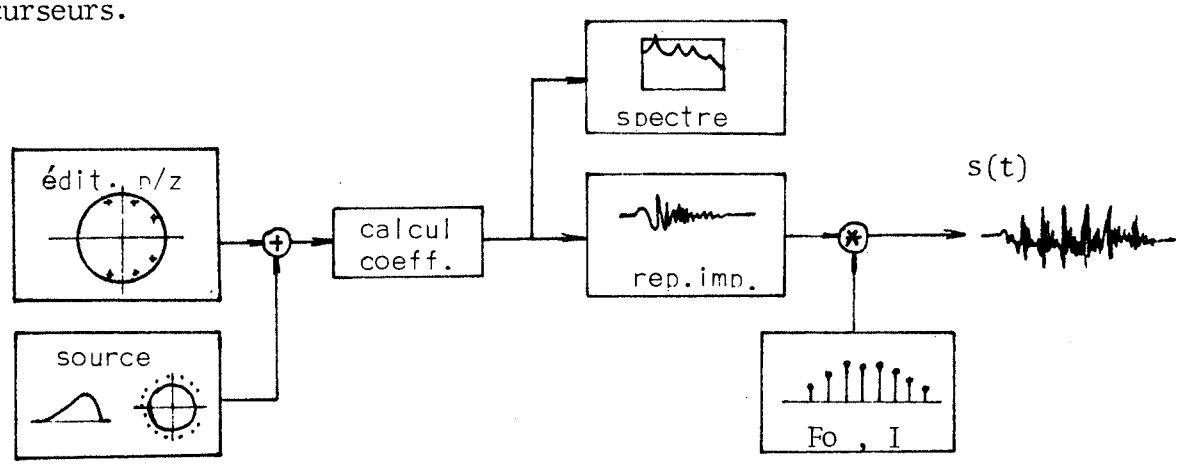


Fig. 1 structure du synthètiseur

Ensuite sont calculés les coefficients de la fonction de transfert, puis la réponse impulsionnelle du conduit vocal. Celle-ci est alors convoluée avec un contour d'intensité et de Fo mesurés par ailleurs.

Au signal ainsi produit, il est possible d'associer le spectre correspondant.

III - CONSIDERATIONS SUR L'EXCITATION

La convolution de la réponse impulsionnelle du conduit vocal par un peigne de DIRAC de pas To ne peut pas produire un son "naturel" : ce modèle simplifié ne

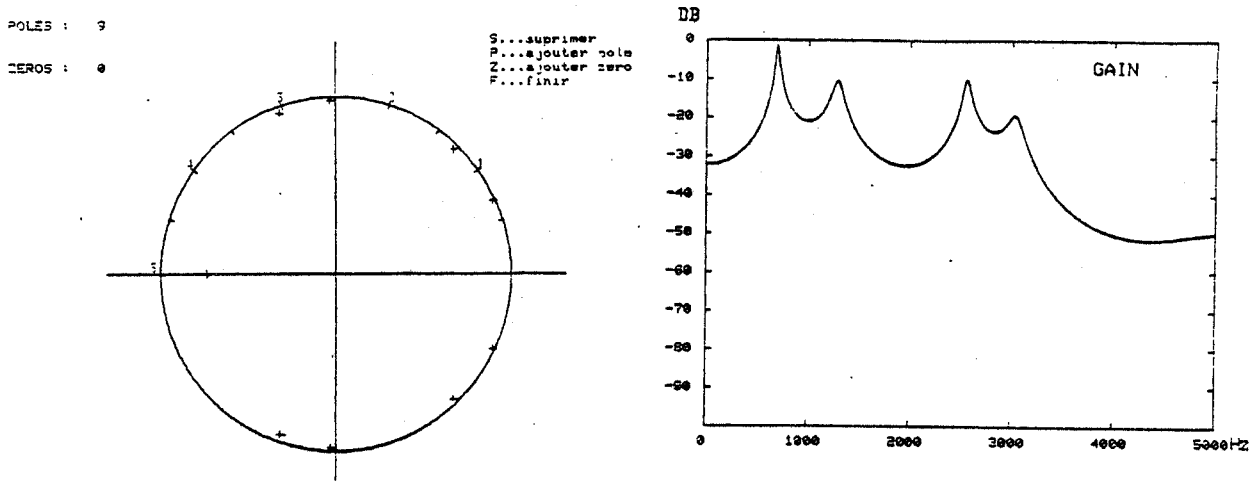


Figure 2 - Répartition des pôles pour la voyelle /a/ et son spectre.

correspond pas à l'excitation d'une source vocale. Ici, contrairement aux synthétiseurs à formants ou à modèles articulatoires, la forme du signal source est fixe. Elle peut être représentée par une série de zéros dans le plan-Z (fig. 3), comme tout signal à durée limitée.

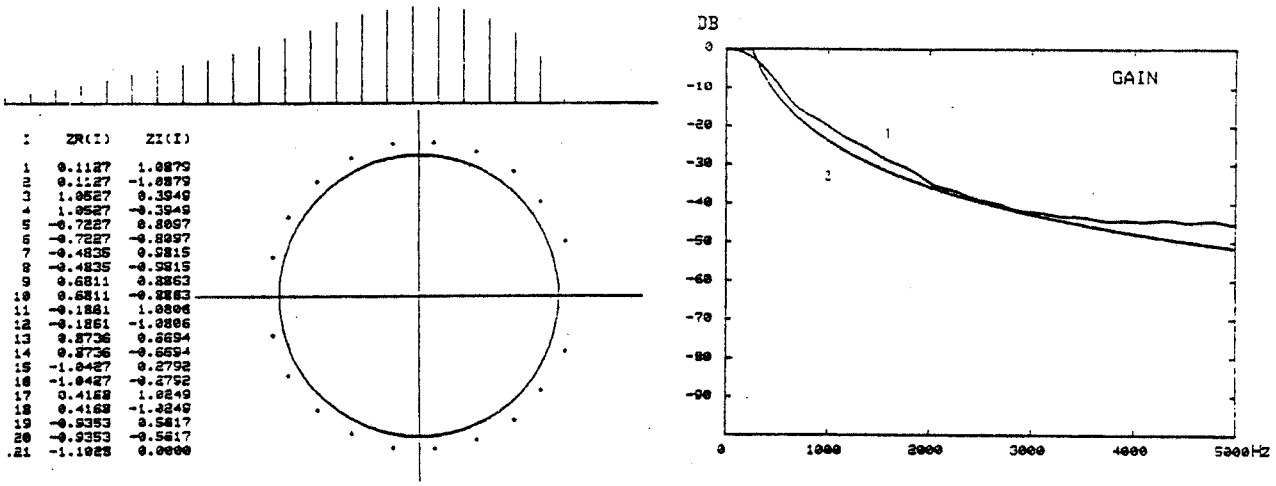
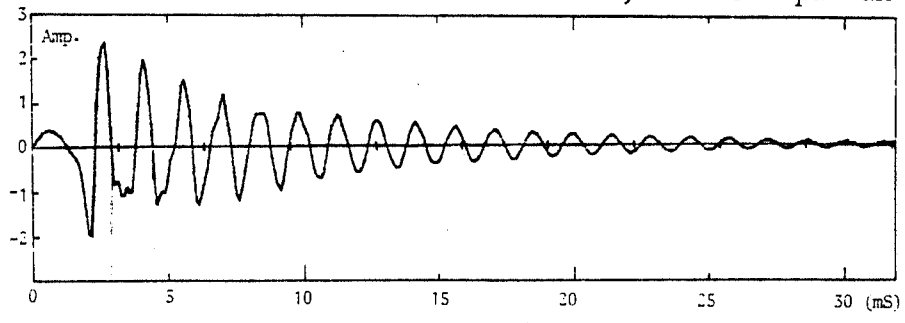


Figure 3 - La fonction d'excitation, ses zéros et son spectre.
(1.Spectre de la source - 2. Référence à -12 dB/octave)

Pratiquement, il suffit de superposer le groupe des zéros relatifs à la source, aux pôles correspondant à la fonction de transfert du conduit vocal, pour rendre compte automatiquement de l'influence de cette source.

En outre, l'adjonction d'un zéro situé à z=1 permet d'approximer convenablement l'effet de la radiation aux lèvres, modélisé par un filtrage à +6 dB/octave.



La figure 4 présente la réponse impulsionnelle correspondant à une voyelle de type /a/.

IV - CONSIDERATIONS SUR L'INTERACTION SOURCE-CONDUIT VOCAL

Sur la figure 4, on peut constater que la réponse impulsionnelle de la source vocale présente un amortissement faible qui provoque, à la convolution, une perduration de son effet sur la période voisine. Pour prévenir ce phénomène, on peut introduire un amortissement plus raide au delà de To. Cependant, un tel système n'est plus linéaire et ses coefficients ne sont plus indépendants du temps. Nous avons opté pour une méthode d'approximation afin de contourner ce problème. Une fenêtre à amortissement exponentiel (fig. 5) vient pondérer la réponse impulsionnelle, stable durant To, puis décroissant avec une constante 0,3 To.

Le signal obtenu à ce terme (fig. 6) correspond à une "parole" relativement naturelle.

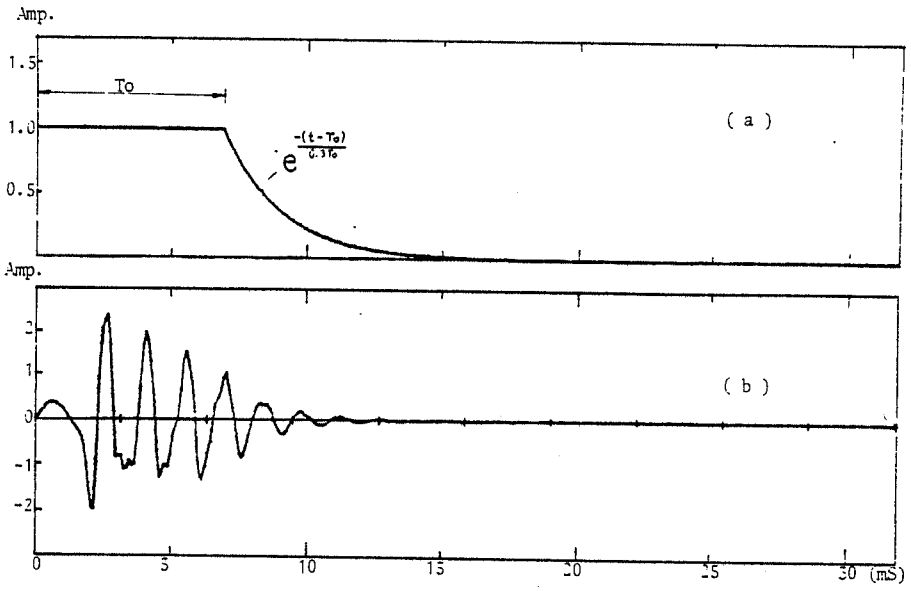


Figure 5 - Fenêtre adoptée pour tenir compte de l'effet de couplage (a). La réponse impulsionnelle correspondante (b).

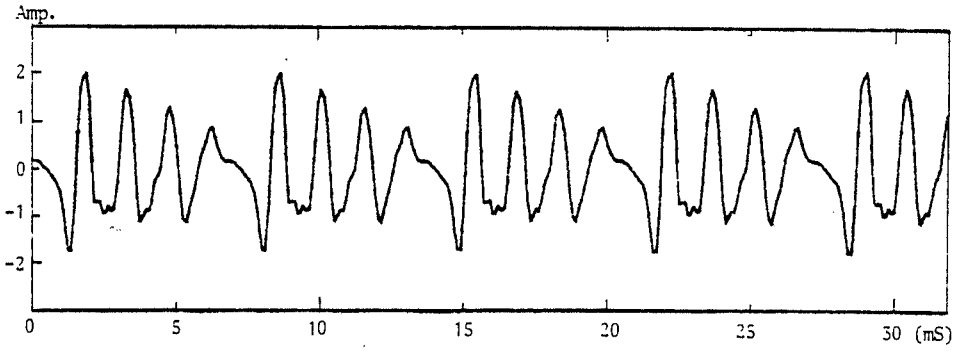


Figure 6 - Voyelle [a] synthétique.

V - CONCLUSION

Nous avons synthétisé l'ensemble des voyelles orales et nasales du français à l'aide de ce synthétiseur. Les résultats obtenus rapidement et de présentation claire, sont satisfaisants. Ils permettent une étude souple de l'influence des pôles et zéros sur la qualité du signal.

Ce synthétiseur présente des liens étroits avec ses homologues à formants : il est possible de lui fournir les paramètres relatifs aux fréquences et bandes passantes.

Il sera modifié pour permettre une évolution temporelle de la structure P/Z.

ETUDE DU CODAGE/DECODAGE DES INFORMATIONS AUDITIVES

J. CAELEN, G. CAELEN-HAUMONT, B. FRAYSSE, H. URGELL

Laboratoire C.E.R.F.I.A. et laboratoire d'Otoneurologie
Université P. Sabatier - TOULOUSE - France.

L'étude du codage des informations auditives apporte à la reconnaissance des ouvertures nouvelles, notamment en fournissant des modèles pour l'analyse acoustique et la recherche d'invariants phonétiques. En réhabilitation des surdités profondes cette étude est indispensable si l'on veut soit adapter la stimulation à chaque cas clinique, soit augmenter la qualité des informations codées en fonction des restes auditifs.

On peut envisager une approche multiple de ce problème :

- par voie physiologique
- par voie psychoacoustique
- en s'aidant de modèles de l'audition

A Toulouse, une équipe pluridisciplinaire s'est regroupée autour de ce thème, comprenant des médecins, orthophonistes, audioprothésistes et scientifiques, autour d'un projet de réhabilitation des surdités profondes par implantation d'électrodes extra-cochléaires - et dans un premier temps d'une mono-électrode.

La voie physiologique :

L'exploration électrophysiologique du tronc cérébral se fait à l'aide de stimuli acoustiques et/ou électriques sur le cobaye. Les réponses des oreilles gauche et droite sont analysées en supposant dans un premier temps que les voies ipsi et contro latérales sont additives. Ceci permet de poser un modèle en terme de théorie des graphes dans lequel les sommets sont les centres nerveux tels que noyaux cochléaires, olives etc... et les arcs sont les réseaux de neurones proprement dits. Chaque sommet possède une fonction de transfert qui lui est propre.

La voie psychoacoustique :

Des signaux sont synthétisés sur ordinateur et envoyés après codage en modulation HF sur l'électrode qui se trouve au contact de la fenêtre ronde. Les premiers résultats semblent indiquer que malgré une perte notable des aigus (relevée par audiogramme) une perception dans cette zone reste possible. La reconnaissance des stimuli est meilleure si l'on ne filtre pas la bande au-dessus de 2 kHz bien que l'audition soit inexistante à l'audiogramme au-dessus. Des listes de mots (paires métathétiques) sont utilisées pour la rééducation phonétique et la mesure des performances des sujets.

La modélisation :

Une approche formelle est en cours d'étude actuellement sur la base des résultats obtenus par voie électrophysiologique. Une approche systématique, expérimentale se fait également en comparant les réponses obtenus par le modèle d'oreille et les réponses obtenues pour certains cas pathologiques. L'ajustement a posteriori des paramètres du modèle, (suppression de cellules par exemple), permet de se rapprocher de cas pathologiques connus. Il est alors possible d'inférer un diagnostic et de prédire la tonotopie des fibres restantes. De telles comparaisons peuvent être effectuées au moyen du potentiel de sommation ou du potentiel microphonique (fig 1 et 2)

Ces premiers travaux confirment la nécessité de regrouper les équipes pluridisciplinaires autour d'un même projet. Les résultats obtenus en réhabilitation montrent d'ores et déjà des performances comparables à celles obtenues à l'aide de multi-électrodes intra-cochléaires.

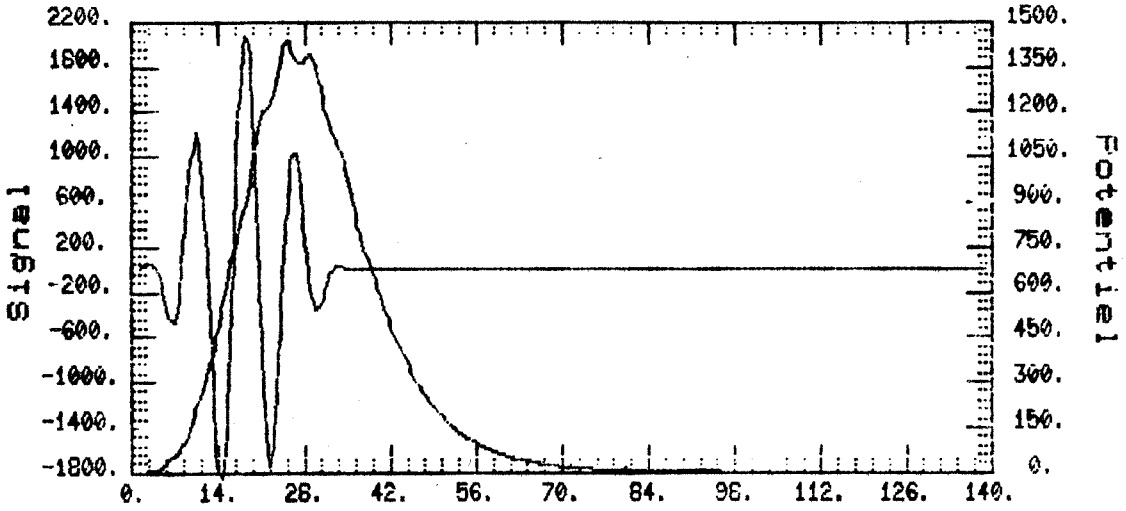


Fig 1 : Stimulus en trait continu, clic filtré à 2 kHz
Potentiel de sommation "normal" obtenu par le modèle d'oreille en grisé

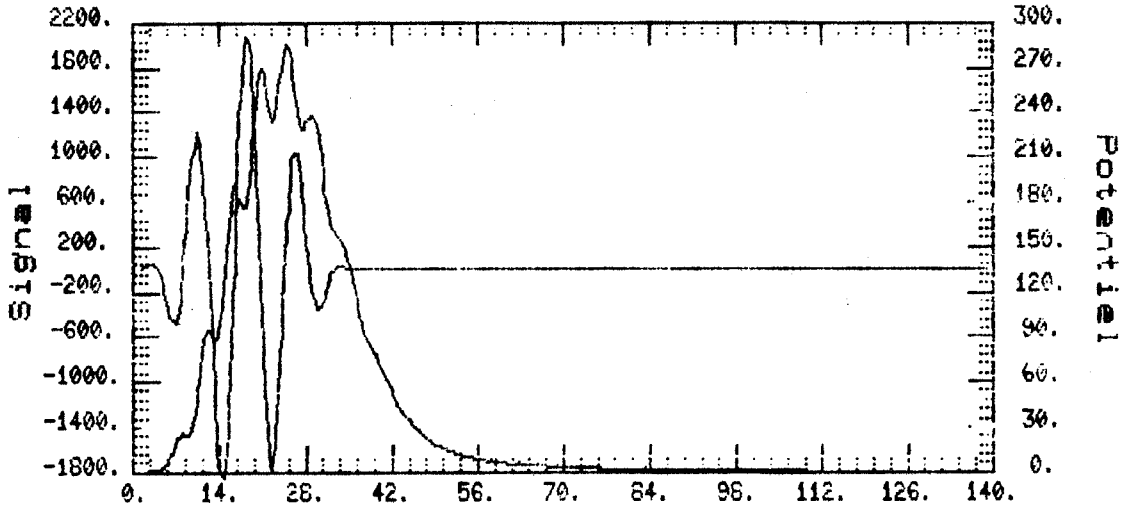


Fig 2 : même légende que pour la fig 1, le potentiel de sommation est obtenu ici en supprimant toutes les fibres nerveuses de fréquence caractéristique supérieures à 1 kHz

DETECTION SYNCHRONE DU PITCH PAR LE CRITERE DE VARIATION D'AMPLITUDE (CVA)

P. JOSPA : Institut de Phonétique - Université Libre de Bruxelles.

De nombreux algorithmes d'extraction du pitch ont pour objet la détection de structures périodiques dans le signal; de ce fait ils sont souvent défaillants lorsqu'ils sont appliqués à des signaux dont la structure périodique est altérée (voix pathologiques, présence de bruit). La méthode du CVA détecte, au contraire, des discontinuités dans le signal liées au fonctionnement de la source laryngée /1, 2/. Cette méthode exploite deux phénomènes bien connus rattachés au mécanisme de la production sonore dans l'appareil vocal: 1) l'instant de fermeture de la glotte est généralement suivi d'un brusque accroissement de l'énergie sonore; 2) entre deux instants de fermeture de la glotte, l'énergie du signal de parole décroît en raison des pertes dans l'appareil vocal. Ces deux phénomènes, qui impriment à l'enveloppe (lissée) du signal d'un son voisé une forme en dent de scie, se manifestent à toutes les fréquences significatives de l'onde de parole. Le CVA fournit, à chaque instant, une estimation de la variation temporelle de l'enveloppe lissée du signal dans une bande de fréquence préalablement choisie. Dans le cas des sons voisés, le CVA présentera donc une succession de maximas prononcés (positifs) aux instants de fermeture de la glotte, et de minimas (négatifs) aux instants de décroissance la plus rapide de l'énergie du signal (fig. 2).La détection synchrone du pitch est dès lors aisée: elle se réduit, pour l'essentiel, à la détection des maximas positifs du CVA dépassant un seuil variable, dynamiquement défini (pics significatifs).

Le principe du calcul du CVA est résumé par le schéma-bloc de la figure 1. Il s'agit essentiellement d'un extracteur d'enveloppe dans la bande de fréquence de 300Hz à 900Hz, suivi d'un lissage; l'enveloppe lissée A(t) est ensuite différenciée: CVA = A(t+Dt) - A(t-Dt), l'intervalle de temps Dt étant nécessairement choisi dans des limites fixées par la largeur de la fenêtre de lissage. Deux versions du CVA ont été implantées. La première est directement calquée sur le schéma-bloc de la fig. 1. La seconde substitue aux deux derniers étages (le lissage suivi de la différence), un seul étage constitué d'un filtre différenciateur passe-bas FIR/3/. Cette seconde version permet de couvrir une gamme plus étendue de fondamentales.

Le choix du filtre différenciateur passe-bas (ou de la fenêtre de lissage) dépend de la hauteur de la voix analysée. En procédant à une analyse par

fenêtres temporelles successives (d'une durée de l'ordre de 50ms), ce choix peut être automatiquement réalisé en se basant sur la dernière valeur de la fondamentale estimée dans la fenêtre précédente.

REFERENCES:

/1/ Jospa P. (1982): R.A. Inst.Phon. Univ. Libre de Bruxelles, no. 17, p.89.

/2/ Jospa P., Schoentgen J. (1982): Fortschritte der Akustik FASE/DAGA '82, Göttingen 1982, Vol. II, p.993.

/3/ Suggéré par Mr Grenez F. du Serv. d'Electr. Gén., Univ. Libre de Bruxelles.

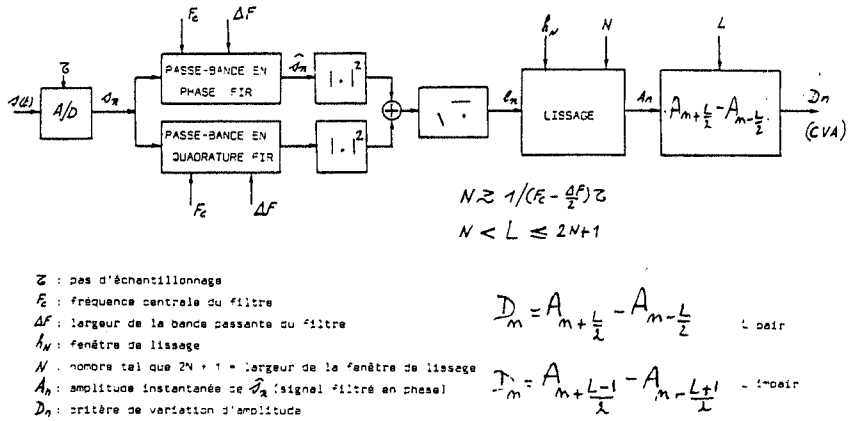


Fig.1

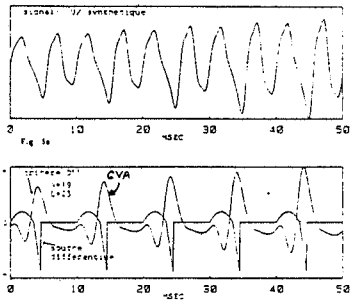


Fig.2

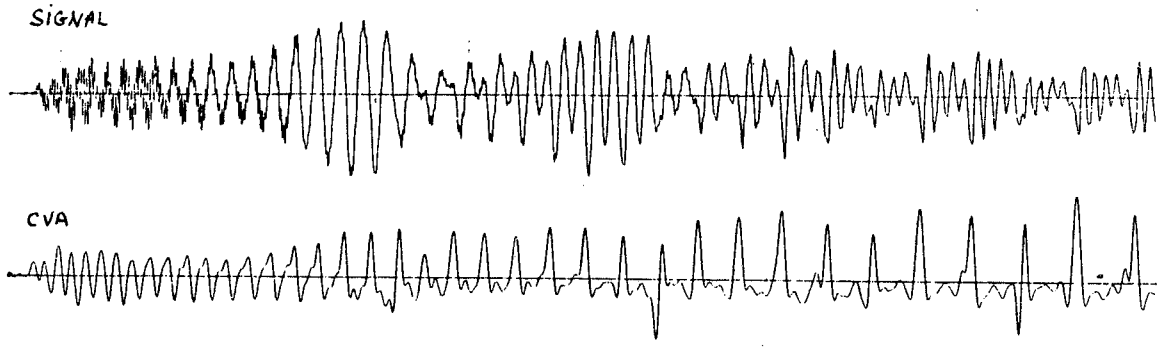


Fig.3

LEGENDES:

Fig.1: Schéma de principe du calcul du CVA.

Fig.2: En haut: voyelle /u/ synthétique. En bas: CVA superposé au signal de source différencié.

Fig.3: En haut: signal synthétique: /i/⟶/a/; la fondamentale varie de 500Hz à 100Hz. En bas: CVA.

Fig.4: En haut: signal naturel (voix féminine): /bo/ extrait du mot /boku/. En bas: le CVA.

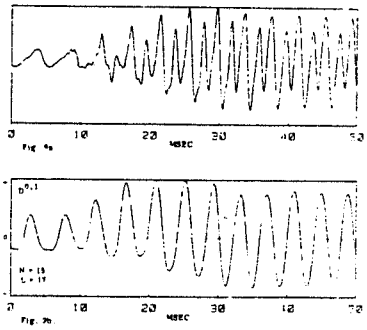


Fig.4

CODAGE DE PAROLE EN TEMPS REEL A DEBIT REDUIT METTANT EN OEUVRE UNE ANALYSE MULTI-IMPULSIONNELLE

J.P. LEFEVRE, O. PASSIEN : Laboratoire Traitement de Parole, CIT-Alcatel
10bis, rue Louis Lormand - 78320 LA VERRIERE

PRESENTATION

Récemment a été introduite l'idée d'exciter, quelque soit la nature des sons, le filtre d'analyse-synthèse LPC classique par une suite d'impulsions convenablement choisie, de telle sorte qu'un critère perceptuel, mesurant la différence entre le son original et le son synthétisé soit minimisé (1). L'amélioration importante de la qualité de la parole synthétique ainsi obtenue a suscité un intérêt considérable pour cette nouvelle approche.

Dans cette communication, nous décrivons un algorithme simple permettant d'obtenir le signal d'excitation du filtre de synthèse, sous la forme d'une suite d'impulsions. La technique présentée ici repose, tout comme la méthode d'analyse par synthèse proposée par Atal et Remde, sur la possibilité de construire un signal estimé, représentatif des impulsions à positionner. Dans notre approche ce signal est obtenu à partir du résiduel et de l'autocorrélation de la réponse impulsionnelle pondérée du filtre de synthèse. Par la suite les impulsions sont positionnées et calibrées une par une. L'utilisation des contraintes de localisation permet un codage effectif de l'information à des débits compris entre 9 et 16 kbit/s. Les résultats de simulations numériques et d'évaluations subjectives nous ont permis de préciser les performances de ce type de codage et nous autorisent à en envisager l'implantation temps réel.

RECHERCHE DES IMPULSIONS

La procédure de recherche des impulsions décrite par ATAL et al. nécessite un grand nombre de calculs (2). Nous rappellerons simplement qu'elle consiste à minimiser l'énergie d'une erreur perceptuelle, pour chaque nouvelle impulsion envisagée.

En fait le coût en calculs peut être considérablement allégé en ramenant la recherche à une simple détection de maximum pour toute nouvelle impulsion.

Tout d'abord, est établi le signal représentatif des impulsions à positionner:

$$t(i) = r(i) + \sum_{j \neq i} (r(i-j) + r(i+j)) * c(j) \; ; \; i = 1,\ldots,N$$

où r(i) représente le résiduel, c(i) l'autocorrélation normalisée et N la longueur de la fenêtre.

Ensuite, les impulsions sont obtenues une par une ; l'impulsion de rang k, d'amplitude a_k et de position i_k étant telle que la condition suivante soit satisfaite :

$$\max_{1 \leq i \leq N} \left| t(i) - \sum_{l=1}^{k-1} a_l * c(|i_l - i|) \right| = k, a_k, i_k$$

Il est clair que la solution d'un tel système ne pose aucun problème.

RESULTATS

L'algorithme proposé ci-dessus a fait l'objet de tests exhaustifs avec des plages de recherche d'impulsions s'étendant sur une fenêtre de 20 ms, combinées avec une analyse LPC d'ordre 14 réalisée à l'aide d'une méthode d'autocorrélation. Typiquement 20 impulsions sont positionnées pour chaque fenêtre. Afin d'obtenir une répartition économiquement codable à débit fixe, certaines contraintes sont imposées sur la position relative des différentes impulsions ; ces contraintes ayant été optimisées de façon à n'entraîner que des dégradations de performance non perceptibles. En allouant 45 à 50 bits pour le codage des coefficients du filtre de synthèse et 7 bits en moyenne par impulsion, un débit fixe de 9600 b/s peut être obtenu.

L'évaluation objective des performances a été réalisée en mettant en oeuvre une distance ceptrale, mesure plus significative pour ce genre de codeur que les traditionnels rapports signal sur bruit. Le dispositif décrit ci-dessus a également fait l'objet d'évaluations subjectives. Pour cela des comparaisons ont été entreprises avec le codage MIC conventionnel (loi A, 8 bits), avec un codage MIC sur 7 bits, et finalement avec le codage MICDA actuellement en cours de normalisation par le CCITT. Les résultats préliminaires obtenus laissent entrevoir pour le codeur LPC avec excitation multi-impulsionnelle décrit dans cet article, une qualité légèrement supérieure au MIC 7 bits et très proche du MICDA-CCITT. Finalement, et tel qu'exposé ci-dessus l'aspect fonctionnement en temps réel a été pour nous une ligne directrice tout au long de la mise au point de cette technique. Actuellement, en simulation (programmation en langage FORTRAN sur VAX/VMS équipé d'un array-processeur FPS-100), le temps d'exécution de l'algorithme est de l'ordre de 0,8 fois le temps réel et nous pensons prochainement atteindre le facteur 0,5.

(1) Atal B.S., Remde J.R., "A new model of LPC excitation for producing natural-sounding speech at low bit rates", ICASSP, Paris, Mai 1982.

(2) Varga A.P. Fallside F., "Multi-pulse excitation in linear predictive synthesis of speech", IEE Colloquium, Londres, N° 1983/31, Avril 1983.

LE POLYMORPHISME ACOUSTIQUE DE /R/ EN FRANCAIS

M. CHAFCOULOFF : Institut de Phonétique, LA 261
AIX-en-PROVENCE - France.

Grâce à l'utilisation de diverses techniques d'investigation physiologique comme la palatographie ou la cinéradiographie, il a été démontré que sur le plan articulatoire, la production du son /R/ était caractérisée par la réalisation de plusieurs variantes allophoniques. Si l'on excepte certains travaux de phonétique comparative, on ne dispose à ce jour que d'un nombre limité d'informations sur les caractéristiques temporelles et spectrales du son /R/. Afin de recueillir des données complémentaires, on a étudié les principaux facteurs responsables de la variabilité acoustique de /R/ en français.

- influence du contexte et de la position.
- interaction source vocale-source de bruit.
- différences inter et intra-locuteurs.

Un corpus de logatomes de type CVCVC a été constitué où la consonne /R/ est analysée en position initiale, intervocalique, finale et en contexte vocalique/i,a,u/.Une procédure expérimentale fondée sur l'examen d'oscillogrammes, de spectrogrammes et l'emploi de différents programmes d'analyse automatique (transformée de FOURIER, prédiction linéaire) a été utilisée.

Les résultats montrent que le son /R/ est d'un point de vue acoustique un phonème essentiellement polymorphe. Le polymorphisme a plusieurs causes.
- Produit à partir d'une source vocale, d'une source de bruit ou d'une combinaison des deux, ses caractéristiques principales seront, en fonction de la prédominance de la source soit celles d'une voyelle (spectre de raies), soit celles d'une consonne (spectre de bruit).
- Particulièrement sensible à la coarticulation, le /R/ subit l'influence de la voyelle contiguë dont la nature conditionne sa réalisation comme allophone vocalique ou consonantique.
- Sa position dans le mot est également déterminante en ce qui concerne sa

réalisation comme allophone voisé ou non voisé. La position initiale est de ce point de vue la plus variable (réalisation de 3 allophones) alors que la position finale est caractérisée par la production d'un allophone unique plus ou moins dévoisé.

D'autre part le son /R/ est également sujet à de nombreuses variations en fonction de l'idiolecte des locuteurs.

- La variabilité intra-locuteur est dépendante du locuteur même et est plus marquée en contexte /a/.
- La variabilité inter-locuteur est généralement importante. L'analyse montre qu'il est possible de relever la présence d'un ou de plusieurs indices acoustiques caractéristiques des réalisations d'un locuteur donné et qui pourraient de ce fait concourir à son identification.

CARACTERISATION PSEUDO-PHONETIQUE DU SIGNAL DE PAROLE

H. MELONI, J. GUIZOL, J. GISPERT: Groupe Intelligence Artificielle
Faculté des Sciences de Luminy - Marseille.

La représentation paramétrique du signal de parole que nous avons choisie diffère sensiblement de celle que nous avions précédemment adoptée (Méloni 82a, 83). L'objectif essentiel de cette nouvelle approche vise le traitement des connaissances acoustico-phonétiques présentées au système sous forme déclarative. Cette technique, partiellement utilisée dans nos travaux antérieurs (Méloni 82b), devrait permettre de collectionner le plus simplement possible les informations (formalisées ou non) disponibles chez les experts phonéticiens. La mise en oeuvre et la stratégie d'utilisation des règles proposées étant effectuées d'une part par la prise en compte de méta-connaissances définies par les spécialistes ou déduites des possibilités particulières de la représentation paramétrique (hiérarchisation à priori des règles par exemples), et, d'autre part au moyen des caractéristiques du signal dans le contexte de l'analyse (valuations d'indices pseudo-phonétiques divers).

La réalisation de ces objectifs impose une représentation des informations acoustiques qui soit suffisamment souple mais proche des caractérisations habituelles des unités segmentales et supra-segmentales manipulées par les spécialistes de la phonétique (traits distinctifs, indices, propriétés, etc..). Cela pose le problème de la nature et de la durée des segments de parole qui constituent les éléments de caractérisation du signal. Nous avons choisi des unités qui représentent des événements pseudo-phonétique (portions stables, variations monotones, variations brutales, etc.). L'expérience nous a montré en effet que les traitements des connaissances phonétiques exigent rarement des informations de niveau plus élémentaire si l'évolution sur le segment de certains paramètres est prise en compte dans le codage. De plus, la définition de ces unités se fait de manière déterministe et utilise des contraintes contextuelles immédiatement disponibles dans le signal. Les pertes d'informations qui résultent d'un tel codage sont largement compensées par l'efficacité des traitements des règles phonétiques.

Un premier codage paramétrique est effectué soit directement sur le signal, soit au moyen d'un spectre lissé obtenu à partir de coefficients de LPC. Les limites des segments sont déterminées essentiellement par les maxima significatifs d'une fonction d'instabilité spectrale. Chaque unité ainsi définie est affectée d'une étiquette qui indique approximativement la nature probable de cette portion de signal (consonantique, vocalique, transition, explosion, constriction, etc.). Ces caractéristiques ne sont pas définitivement acquises mais constituent des repères utiles pour les stratégies d'utilisation des règles phonétiques.

Chaque segment ainsi défini est caractérisé par de nouveaux paramètres dont certains pourraient être assimilés à des indices acoustiques valués sur des échelles variables (entre 2 et 20 positions). Les attributs retenus sont de nature variée, certains nécessitent un ajustement contextuel pour en extraire l'information pertinente (durée, énergies, FO, etc.), tandis que d'autres sont normalisés (amplitudes comparées des pics, émergence du 1er pic, etc.) ou immédiatement utilisables (position des pics, rapports des énergies, etc.). Ils représentent souvent une valeur moyenne sur un segment et parfois l'évolution temporelle d'une mesure sur la portion de signal correspondante. L'évaluation moyenne est effectuée tantôt sur la zone de plus grande stabilité (pics des segments vocaliques, énergies, etc.) tantôt sur la portion de plus grande variation (explosions par exemple). De plus, chaque spécification est exploitée de diverses manières

pour déterminer les traits particuliers des phonèmes et des accents.

Les paramètres et indices se répartissent schématiquement dans les classes suivantes : 1) paramètres indépendants du spectre, 2) répartition de l'énergie dans le spectre (indépendamment des pics), 3) caractérisation des pics spectraux, 4) variations temporelles de paramètres sur le segment. A titre d'exemple, si nous souhaitons étudier le caractère compact d'une portion de signal correspondant à une voyelle, nous disposons sur les segments d'indices valués qui expriment les propriétés suivantes :

- position dans le spectre de la fenêtre (de largeur 800 Hz) dans laquelle l'énergie est maximale,
- rapport de l'énergie maximale à l'énergie utile du spectre (80-3600 Hz),
- distance entre les centres de gravités des portions de spectre situées en deçà et au delà de 1500 Hz (valeur approximative qui peut être ajustée à chaque locuteur),
- distance entre les pics P1 et P2 et amplitudes normalisées de ces maximas, etc.

Les unités phonétiques (phonèmes et accents) que nous souhaitons décrire peuvent être caractérisées par un nombre variable et limité de paramètres (tous ne sont pas pertinents pour une unité donnée). Ainsi, la projection sur un ensemble de paramètres d'un phonème quelconque doit tenir compte du locuteur, des éléments les plus résistants aux variations contextuelles ainsi que de leur pouvoir de discrimination dans un environnement particulier.

REFERENCES

(GUIZOL 84) Guizol J., Méloni H. "apprentissage des règles d'interprétation d'événements pseudo-phonétiques" 13èmes JEP, Bruxelles 1984.

(MELONI 82a) Méloni H. "Etude et réalisation d'un système de reconnaissance automatique de la parole continue", Thèse d'Etat, Université d'Aix-Marseille 2, Février 1982.

(MELONI 82b) Méloni H., Guizol J. "Utilisation de paramètres prosodiques dans un système de reconnaissance de la parole continue", Séminaire Prosodie et Reconnaissance, Aix en Provence, Octobre 1982.

(MELONI 83) Méloni H., Guizol J. "Identification d'événements pseudo-phonétiques pour la reconnaissance automatique de la parole continue", 11ème ICA, Toulouse, Juillet 1983.

UN PERIPHERIQUE D'ENTREE-SORTIE ANALOGIQUES A GRANDE VITESSE ET HAUTE RESOLUTION

B. TESTON : Institut de Phonétique, L.A. 261 C.N.R.S.
Université de Provence - AIX-en-PROVENCE

Objet: Entrée: Digitalisation de signaux analogiques dans la bande passante auditive (essentiellement signaux de parole) et, signaux de paramètres physiologiques tels que par exemple, électomiographiques et aérodynamiques.

Sortie: Synthèse vocale au moyen de différents procédés, génération de stimulus auditifs pour des experiences de perception, en particulier pour l'audiomètrie dichotique.

Caractéristiques: Entrée: Signal codé sur 8, 10, 12, 14, 16 bits au choix.
Dynamique de 90 dB avec un niveau d'entrée de + ou - 10 Volts et 305 microVolts de résolution.
Gain de 60 dB par bonds de 20 dB (signal max de + ou - 10 milliVolts) sur entrée unipolaire de 10 kOhm d'impédance, et de 100 dB par bonds de 20 dB (signal max de + ou - 100 microVolts) sur entrée différentielle de 1 kOhm d'impédance.
Convertisseur: Burr-Brown PCM 75 KG.
Echantillonneur bloqueur: Analogic MP 201 A.
Filtres: Passe bas commutable de 24 dB de chute par octave à partir de 80 Hz.
Réjecteur à 50 Hz également commutable.
Antirepliement: Passe bas à fréquence de coupure variable élliptique d'ordre 6 (à capacités commutées Reticon R5609), à fréquence fixe de 15 kHz Rifa PBA 3179.
Cadence d'échantillonnage maximale: 60 kHz avec 16 bits de résolution.
Distortion (THD): Inférieure à 0,1% de 20 à 15 kHz pour un signal d'entrée de + ou - 1 Volt avec le filtre Rifa.
Bruit: Inférieur à -95 dB dans les mèmes conditions.
Le départ de l'acquisition peut être déclenché par programme, sur un seuil variable du signal ou manuellement.

Sortie: Signal codé sur 16 bits.
Dynamique de 90 dB avec un niveau de sortie de + ou - 10 Volts et 305 microVolts de résolution.
Niveau de sortie: + ou- 10 volts sous une impédance de 50 Ohm, et + ou - 1 Volt sous une impédance de 1 kOhm.
Convertisseur: Burr-Brown PCM 50 KG.

Atténuateur de distorsion(deglitching):Analogic MP 201 A.
Filtre antirepliement:A fréquence de coupure variable élliptique d'ordre 6 (à capacités commutées Reticon R5609),à fréquence fixe de 15 kHz Rifa PBA 3179.
Il existe deux voies de sortie identiques,le démultiplexage est effectué sur les données numériques.
Cadence d'échantillonnage maximale:60 kHz sur une voie,30 kHz par voie sur deux voies.Les deux voies peuvent être utilisées alternativement sur deux postes de travail.Les échantillons apparaissent en même temps sur les deux voies en utilisation simultanée(stéréophonique ou dichotique).
Distorsion (THD):Inférieure à 0,1% de 20 à 15 kHz pour un signal de sortie de + ou - 10 Volts sous 50 Ohm d'impédance avec le filtre Rifa.
Bruit:Inférieur à -100 dB dans les mèmes conditions.

UTILISATION:Le périphérique est connecté à un calculateur Digital PDP 11-24 par l'intermédiaire d'une carte d'entrée-sortie DRU11 qui gère automatiquement le basculement des tampons en accés direct mémoire.La cadence d'échange est limitée à 60 kHz sur notre système par la vitesse de lecture-écriture des disques RL02.
Le pilote du périphérique est géré sous le système d'exploitation UNIX (1).
Le cadencement de l'acquisition et de la restitution est réalisé par une horloge à quartz.Il est possible de télécommander des magnétophones pour faire des acquisitions et des bandes de tests automatiquement.
Le périphérique est isolé electriquement du calculateur par des coupleurs optiques.

(1)ESPESSER.R,"Gestion d'un convertisseur numérique-analogique sous UNIX", Actes du séminaire GALF,Paris,15-16/12/1983,155-163.

DETERMINATION EXACTE DU FONDAMENTAL AVEC UN LARYNGOGRAPHE

W. HESS et H. INDEFREY : Lehrstuhl für Datenverarbeitung, Technische Universität München - Postfach 202420, D-8000 MÜNCHEN 2, R.F.A.

Pour plusieures raisons bien connues la détermination du fondamental compte parmi les problèmes les plus délicats dans le domaine d'analyse de la parole. Une multitude d'analyseurs et algorithmes ont été developpés pour cette tâche (Hess, 1983). Néanmoins, il n'y a aucun analyseur qui fonctionne sans erreur.

Afin d'évaluer le fonctionnement d'un analyseur, comme il a été fait, par exemple, dans l'étude classique par Rabiner et al. (1976), il faut qu'on établisse une base de données qui contient d'information de référence sur le fondamental. Si l'on crée cette base de données manuellement ou avec un système interactif, cela exige beaucoup de travail humain (Rabiner et al., 1976). On ne peut générer les données automatiquement qu'en se servant d'un instrument qui mesure le signal excitateur du larynx plus directement que les analyseurs du signal parole. Pour cette raison nous avons décidé d'utiliser un laryngographe (Fourcin et Abberton, 1971) qui mesure les fluctuations de l'impedance laryngienne pendant les cycles glottiques. Quand la glotte se ferme, l'impedance laryngienne tombe brusquement, ce qui produit une discontinuité dans le laryngogramme (Fig.1b), c'est-à-dire, un maximum signifiant dans la dérivée du laryngogramme par rapport au temps (Fig.1c). Ce maximum coincide avec l'instant d'excitation maximale de conduit vocal; il est donc bien approprié à servir comme référence pour le début d'une période fondamentale. Notre algorithme détermine ces maxima dans la dérivée du laryngogramme; des pics parasitaires dûs au bruit étant supprimés par une analyse simple de seuil. Une interaction manuelle peut être limitée á des signaux irréguliers ou à des cas où un pic individuel n'est pas sûrement accepté ou rejeté. Avec cette base de données qui aussi contient le signal parole enregistré simultanément avec le laryngogramme, on peut tester n'importe quel analyseur du fondamental et particulièrement ces analyseurs qui opèrent dans le domaine temporel et mesurent le signal période par période.

Bien que l'usage du laryngographe assure que l'algorithme ne fasse plus de grosses erreurs (tant que le laryngographe ne faillit pas pour un locuteur individuel), on y trouve encore des petites erreurs de quantification dûs à l'échantillonnage du signal. Comme les données doivent satisfaire à des exigences quelconques par rapport à l'exactitude et la qualité du mesurage, et comme le seuil différentiel de l'oreille par rapport à des changes du fondamental (Flanagan et Saslow, 1958) représente l'exigence la plus sevère, il faut que les erreurs de quantification n'excèdent pas une valeur de 0,5% sur toute la gamme de F_0 (50-500Hz). Cette demande requiert donc une fréquence d'échantillonnage de 100kHz au moins. Il n'est réaliste ni d'échantillonner le signal analogique à une telle

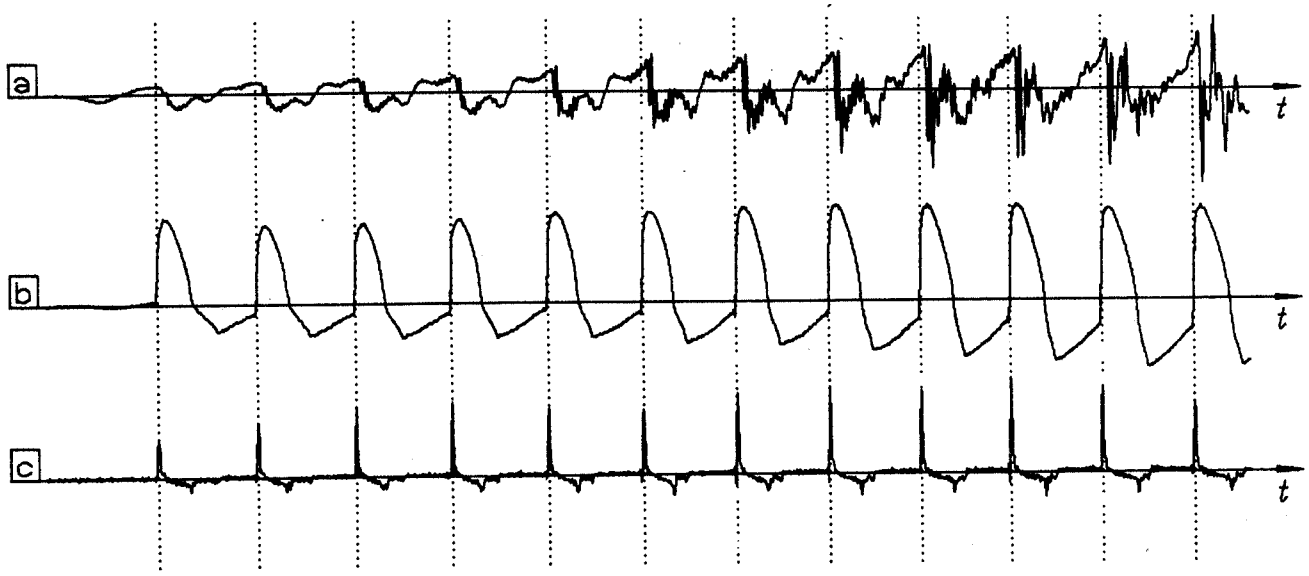


Fig.1a-c. Signal parole (**a**), laryngogramme (**b**) et dérivée du laryngogramme par rapport au temps (**c**).

fréquence ni d'augmenter la fréquence d'échantillonnage avec un filtre numérique pour tout le signal. Notre algorithme détermine donc les maxima signifiants dans le laryngogramme échantillonné à la frequence utilisée pendant l'enregistrement (16 kHz); puis il augmente cette fréquence à 128 kHz dans un petit intervalle (500 µs environs) autour de chaque maximum. Le filtre numérique passe-bas employé pour cette tâche, étant d'un ordre de 144, a été developpé en utilisant l'algorithme de Parks et McClellan (1972). Afin de ne pas changer la position des pics, le filtre est réalisé dans une version non causale, c'est-à-dire, avec une réponse de phase égale à zéro. Comme le filtre ne calcule que très peu d'échantillons à cette fréquence augmentée, la dépense de calcul pour cette interpolation n'est pas grave.

Les données fournies par cet algorithme sont libres de grosses erreurs du fondamental; l'erreur de quantification est moins que 0,5% sur toute la gamme de F_0. Elles peuvent donc servir comme données de référence pour l'évaluation comparative des analyseurs du fondamental.

Bibliographie

Flanagan J.L., Saslow M.G. (1958): "Pitch discrimination for synthetic vowels." Journ. Acoust. Soc. Am. **30**, 435-442

Fourcin A.J., Abberton E. (1971): "First applications of a new Laryngograph." Med. Biol. Illustr. **21**, 172-182

Hess W. (1983): **Pitch determination of speech signals - algorithms and devices** (Springer, Berlin)

Hess W., Indefrey H. (1984): "Accurate pitch determination of speech signals by means of a laryngograph." Proc. IEEE ICASSP-84, San Diego (IEEE, New York)

Parks T.W., McClellan J.H. (1972): "Chebyshev approximation for nonrecursive digital filters with zero phase." IEEE Trans. CT-**19**, 189-194

Rabiner L.R., Cheng M., Rosenberg A.E., McGonegal C.A. (1976): "A comparative study of several pitch detection algorithms." IEEE Trans. ASSP-**24**, 399-413

UN LOGICIEL DE TRAITEMENT DU SIGNAL DE PAROLE SOUS UNIX

R. ESPESSER : Institut de Phonétique, LA 261 CNRS,
29, av. R. Schumann, 13621 AIX-en-PROVENCE, France.

Nous décrirons essentiellement ce qui est spécifique du système d'exploitation (UNIX V7) supportant ce logiciel, par ailleurs classique quant à ses possibilités de traitement et d'édition de signal. Il est constitué de modules indépendants, appelés fonctions, chacune étant un processus directement exécutable sous le contrôle de l'interpréteur de commande d'UNIX (le "shell"); vu ses possibilités, une interface plus spécifique nous a paru inutile, voire limitative. Les fonctions sont de deux type:

-fonction "filtre": tous les modules de traitement sont des "filtres", traitant les données présentes sur leur entrée standard (ES) tant qu'elle est non vide, écrivant les résultats sur leur sortie standard (SS);(pour les fonctions à "entrée sortie multiple", -analyse (synthèse) LPC par ex.- l'usager choisit entre une sortie (entrée) sur fichiers distincts et/ou une sortie(entrée) formatée sur la SS (ES)). Les filtres sont par définition indépendants du flux principal d'E/S, et en particulier n'ont pas à gérer l'accès au signal d'entrée: support, position et durée.

-fonctions accédant directement au fichier spécial(FS) au sens d'UNIX, supportant le signal numérisé, principalement des fonctions d'édition: recherche interactive, interfaces entre le FS et l'exterieur(ex: "l" copie FS sur SS), conversion numérique/analogique, gérée par un nouvel appel-système(1), gestion de l'offset ou pointeur courant(PCO) dans le FS, gestion de marqueurs symboliques sur le FS,etc..

A chaque fonction est associé un fichier(le descripteur) contenant les valeurs courantes paramètrant le traitement, eventuellement modifiées lors de la commande; à un instant donné, l'ensemble des descripteurs(et de certains fichiers de travail) constitue le contexte de l'usager, normalement situé dans son répertoire courant.

Les fonctions filtres accèdent aux échantillons par spécification d'un fichier d'entrée ou par connexion à la fonction "l" via un "pipe"(tube,"dataduc" ?)

ex: modul < echant > resul la fonction "modul" traite la totalité du fichier "echant", les resultats sont conservés dans le fichier "resul".

ex: l 200 | modul > resul le même traitement ne porte que sur 200 ms dans le FS à partir du PCO. L'usager peut ainsi combiner immédiatement

des fonctions classiques, les siennes propres et des utilitaires d'UNIX; ainsi
1 2000 | vnv | tee vois | f01 | graph > imag effectue à partir du PCO sur une durée de 2s sur le FS, une détection de voisement, sauvegardée au passage dans "vois" (utilitaire "tee"), une détection de F0, un tracé de courbe différé et conservé dans "imag"; la commande suivante:
f02 < vois | graph | cat - imag effectue une autre détection de F0 à partir du fichier "vois" précedent, la préparation de la courbe correspondante, et le tracé effectif des deux courbes par concaténation de "imag" et de l'ES de l'utilitaire "cat".

UNIX permet facilement l'exécution de processus en parallèle, par exemple exécuter une analyse spectrale et simultanément repérer le prochain segment à traiter; d'où des conflits d'accès aux descripteurs, résolus par l'emploi de la notion d'environnement d'un processus: le contexte courant de l'usager est copié dans un répertoire temporaire dont le nom est rangé dans une variable du "shell"; cette variable est interne à la procédure lancée en parallèle et fait partie de son environnement; les modules d'ouverture des descripteurs connaissant cette variable, les fonctions travaillent ainsi dans leur contexte privé; la procédure "pars" gère l'ensemble pour l'usager, par exemple:
pars "1 2000 | vnv | tee vois | f01 | graph > imag" effectue en parallèle un traitement déja vu, l'usager devant seulement ne pas modifier la zone du FS concernée, mais pouvant, sous cette réserve, exécuter d'autres traitements, via "pars" éventuellement.

La structure adoptée pour ce logiciel de traitement nous parait constituer un système ouvert, de maintenance très aisée, permettant un style varié de commande, allant de la simplicité de procédures figées regroupant des combinaisons statiques de fonctions à la souplesse plus complexe des exemples précédents.

(1) Gestion d'un convertisseur numérique/analogique sous Unix, R. ESPESSER, séminaire GALF-GRECO analyse du signal de parole, Paris 15-16 Déc. 1983

ALIGNEMENT DES FRONTIERES PHONEMIQUES SUR LE SIGNAL

M. BREANT et J. CAELEN : CERFIA, Université Paul Sabatier
TOULOUSE - France.

Nous présentons un système de segmentation automatique en phonèmes d'une phrase donnée. En entrée, on dispose :1-de la transcription phonémique exacte de la phrase prononcée par le locuteur ; 2-du signal acoustique et de paramètres calculés en tout point du signal. En sortie, on obtient sur le signal toutes les frontières phonémiques.

Le traitement se fait en 2 parties. On détermine d'abord les pauses sur le signal puis on traite les chaînes phonémiques inscrites entre 2 pauses par un algorithme de programmation dynamique (du type "Sakoe et Chiba")dont les sorties sont corrigées par un algorithme de postcorrection.

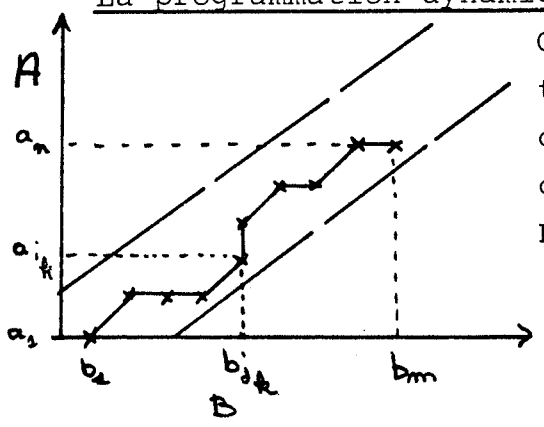
Le traitement des pauses : les pauses du signal sont des longs segments étiquetés "occlusif soud" parmi lesquels on élimine les occlusives de fin de groupe et les occlusives doubles. Une bijection ordonnée affecte alors une pause-phonème à chaque pause-signal.

La transformation chaîne phonémique/chaîne acoustique : toute chaîne phonémique est transformée en chaîne acoustique. On découpe les phonèmes non homogènes en éléments acoustiques homogènes. Ex : décomposition de /p/ en fin de groupe :

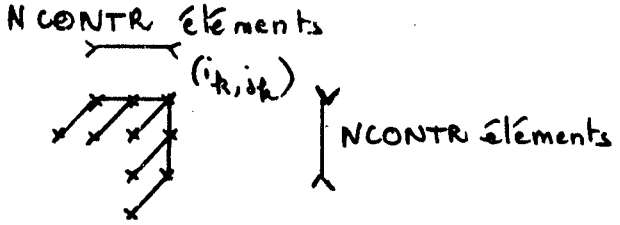
/p/ ⟶ implosion occlusion explosion friction

La présegmentation du signal : On découpe le signal en segments (infraphonémiques) égaux en longueur. Celle-ci est fonction du débit du locuteur.

La programmation dynamique :



On recherche le chemin optimal permettant de projeter $A=(a_i)_{1\leq i\leq n}$, le signal découpé en n segments sur une chaîne d'éléments acoustiques de référence $B=(b_j)_{1\leq j\leq m}$



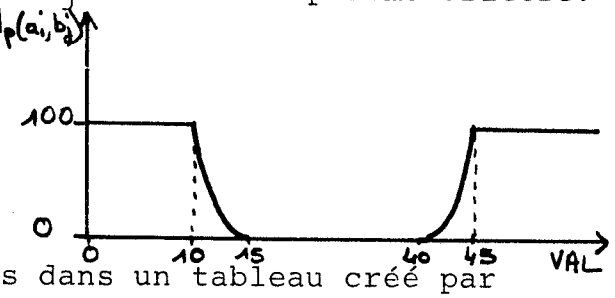
D'autre part, on utilise une contrainte de pente (NCONTR) pour forcer le chemin à ne pas s'écarter de la diagonale. Si NCONTR=3 (voir ci-dessus), (i_k, j_k)ne peut avoir que les antécédents présentés ci-dessus. Les chemins autorisés sont pondérés par des coefficients, fonctions des éléments acoustiques.

Le calcul de distance i.e. le calcul de $d(a_i, b_j)$ mesure de dissemblance entre a_i et b_j: A chaque a_i est attaché un ensemble de critères qu'il doit vérifier, à chaque critère est attaché un paramètre à valeur scalaire ou non. KCRITER

$$d(a_i, b_j) = \frac{1}{\text{KCRITER}} \sum_{p=1} d_p(a_i, b_j)$$

KCRITER est le nombre de critères attaché à b_j. $d_p(a_i, b_j)$ est la mesure de dissemblance entre a_i et b_j au sens du p-ième critère.

Exemple : soit b_j=/m/ et soit p le critère énergie. Le calcul de $dp(a_i, b_j)$ est figuré ci-contre ; VAL est la valeur moyenne en dB de l'énergie sur le segment a_i. Les valeurs 10,15,40,45 sont répertoriées dans un tableau créé par l'expérience.



L'algorithme de post correction : Il a pour but d'affecter à une zone de parole un et un seul élément acoustique (certains éléments sont éliminés sur une zone de parole, certaines zones sont redécoupées). Les zones correspondant à plusieurs éléments acoustiques d'un même phonème sont regroupées en une seule. D'autre part, les positions des frontières des consonnes vocaliques intervocaliques sont améliorées.

UTILISATION D'UN SYSTEME DE RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEUR SUR UN AUTOCOMMUTATEUR PRIVE

B. FLOCON, J. SAP : Laboratoires de Marcoussis - Centre de Recherches de la C.G.E.
Route de Nozay - 91460 MARCOUSSIS - France.

Le système de reconnaissance de mots isolés multilocuteur qui est utilisé est le système SYRIL [1] : le système dispose d'un vocabulaire de 35 mots, comportant les 10 chiffres, les constituants de nombres, les quatre opérateurs et quelques mots de fonction. Le système fonctionne sur des données calculées toutes les 16 millisecondes; il s'agit de 9 coefficients du cepstre calculés selon une échelle de fréquences Mel [2] et d'un terme représentant l'énergie du signal [1]. L'ensemble de références a été obtenu à l'aide d'une méthode de classification automatique appliquée à un ensemble de 60 représentants de chaque mot du vocabulaire provenant de 20 locuteurs différents à travers un microphone de bonne qualité (SHURE SM10).

Le but de cette étude est de montrer que l'on peut utiliser l'ensemble de références obtenue à partir du microphone pour une utilisation en reconnaissance à partir du combiné téléphonique sur un autocommutateur privé. Il suffit pour cela de constituer un ensemble d'apprentissage de dimensions très réduites à partir du combiné téléphonique : 4 locuteurs ont enregistré simultanément les 35 mots à travers d'une part le microphone de bonne qualité et d'autre part à travers le microphone du combiné téléphonique.

La comparaison des signaux provenant des 2 microphones permet de déterminer une fonction de transfert à appliquer à l'ensemble de références initial afin de pouvoir l'utiliser à partir du combiné téléphonique. Deux types de fonction de transfert ont été déterminés : une fonction de transfert indépendante du temps, obtenue en déterminant une moyenne des différences entre le mot "microphone" et le mot "combiné" sur l'ensemble du mot, et une fonction de transfert dépendante du temps, obtenue grâce à la création d'un mot "différence", créé à partir des différences à chaque instant entre le mot "microphone" et le mot "combiné".

Les travaux que nous avons réalisés ont comporté trois étapes : détermination d'une fonction de transfert, puis création d'un nouvel ensemble de références en ajoutant la fonction de transfert à l'ensemble initial, et enfin test de reconnaissance en utilisant le nouvel ensemble de références.

Résultats : un ensemble de 12 locuteurs-test ne faisant pas partie des deux précédents groupes de locuteurs a été constitué. Chaque locuteur a prononcé une fois la liste des 35 mots à partir du combiné téléphonique de son bureau. Les performances du système de reconnaissance utilisant l'ensemble de références initial sont de 80 % de mots bien reconnus. Lorsqu'on détermine un ensemble de références en utilisant la fonction de transfert dépendant du temps, elles passent à 88 %, avec la fonction de transfert indépendante du temps nous obtenons 91 % de bonne reconnaissance. Ces chiffres sont à comparer aux performances de SYRIL (reconnaissance à partir du microphone de bonne qualité) [1] qui sont de 98 %.

[1]"SYRIL : Système temps réel de reconnaissance de mots indépendant du locuteur". B. Flocon, N. Briant. 4ème congrès AFCET Intelligence Artificielle et Reconnaissance des Formes, Paris, Janvier 1984.

[2]"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" , S. B. Davis, P. Mermelstein, IEEE Trans. on ASSP ASSP-28 numero 4, Août 1980.

LA RECONNAISSANCE DES OCCLUSIVES SOURDES EN FRANCAIS

G. MERCIER : LAA/TSS/RCP - C.N.E.T. - Route de Trégastel
22301 LANNION Cedex - France.

INTRODUCTION

Nous décrivons ici un algorithme, en cours de mise au point, permettant de détecter les occlusives et de reconnaître leur lieu d'articulation. Ce module intégré dans le programme d'analyse phonétique de "KEAL" comprend 4 étapes principales : la détection de l'occlusion et de l'explosion, la séparation entre occlusives sourdes et occlusives voisées, la distinction entre l'explosion d'une occlusive et l'attaque d'une voyelle et la reconnaissance du lieu d'articulation ; la deuxième étape ne sera pas traitée dans cette communication.

DETECTION DES OCCLUSIVES

Les occlusives sont caractérisées par la présence d'une zone de stabilité correspondant soit à la barre de voisement soit au silence précédant l'explosion suivie d'une zone de forte instabilité correspondant à l'explosion et à la transition vers la voyelle.

La zone stable de l'occlusion est caractérisée par les indices suivants : énergies et centres de gravité des spectres peu élevés. La zone instable qui suit est caractérisée par une très forte variabilité spectrale (distance entre deux spectres consécutifs calculés à 13.3 ms d'intervalle) et par la présence d'un maximum d'énergie dans les 26 ms suivant le début de l'explosion.

DISTINCTION OCCLUSIVE-ATTAQUE DE VOYELLE, APRES UN SILENCE

Il est parfois difficile de distinguer après un silence l'attaque d'une voyelle non précédée d'une consonne, du spectre de bruit de l'explosion d'une occlusive sourde. Pour identifier une occlusive, on a recours aux indices suivants :

- La dérivée et l'énergie du spectre passent par un minimum très prononcé pendant la zone instable.

- L'écart entre le centre de gravité de la voyelle et celui du spectre moyen de la zone instable est important.
  Pour /k/ et /t/, ce centre de gravité passe par un maximum très prononcé pendant l'explosion.

RECONNAISSANCE DU LIEU D'ARTICULATION DES OCCLUSIVES SOURDES

Pour le moment on utilise 3 indices calculés sur la zone instable dont on a préalablement déterminé les frontières.

- on calcule le spectre moyen sur cette zone ainsi que la position (centre de gravité en fréquence) et l'amplitude des maxima d'énergie dans les 3 bandes de fréquences suivantes : [250-850 Hz] , [850-2800 Hz] et [2800-4200 Hz] . Ces amplitudes respectivement notées : Maxbf, Maxcf et Maxhf sont comparées ; suivant que l'amplitude maximale correspond à maxbf, maxcf ou maxhf on renforcera respectivement la probabilité de /p/, /k/ ou /t/.
- on observe aussi l'évolution du centre de gravité dans la zone des basses fréquences [250-850 Hz] : si ce centre de gravité croît du début à la fin de l'explosion c'est probablement /p/ sinon c'est /t/ ou /k/.
- s'il n'existe pas de minimum ni de palier sur la courbe représentant les variations de l'énergie du début à la fin de la zone instable et si l'énergie dans la bande 250-450 Hz est relativement élevée dès le début de l'explosion, on renforce la probabilité de /p/.

RESULTATS LT CONCLUSION

Le taux actuel de reconnaissance du lieu d'articulation des occlusives sourdes est de 70 % pour la parole continue (7 locuteurs masculins, 320 occurences extraites de phrases phonétiquement équilibrées) ; ce score s'élève au-dessus de 80 % pour les mots isolés. Cette méthode est malgré tout encore trop simple et la prise en compte du contexte devrait améliorer les résultats ainsi que l'indiquent d'autres études (DE MORI, TANAKA, BONNEAU, DELGUTTE).

RECONNAISSANCE ANALYTIQUE MULTILOCUTEURS DE MOTS ISOLES

L.C. SAUTER, D. GROSSETETE-FOURNOL : Laboratoires de Marcoussis
Centre de Recherches de la CGE - Route de Nozay
91460 MARCOUSSIS - France.

Les systèmes de reconnaissance multilocuteurs de mots isolés utilisent en général plusieurs exemples de chaque mot du vocabulaire prononcés par un nombre assez grand de locuteurs différents. L'apprentissage d'un vocabulaire représente alors un travail considérable. Pour cette raison, nous étudions un système de reconnaissance analytique de mots isolés indépendant du locuteur. Chaque mot du vocabulaire pourra être représenté par un treillis de segments de référence et la reconnaissance d'un mot sera le choix du treillis minimalisant une certaine distance. Chaque segment sera représenté par un certain nombre d'empreintes extraites d'élocutions de plusieurs locuteurs. Différentes approches sont possibles pour chercher le mot le plus proche, selon qu'on effectue ou non une segmentation préalable, selon le caractère ascendant ou descendant de l'algorithme de recherche de l'optimum.

1. Sans segmentation

a) Reconstitution d'empreintes globales : les différentes variantes de chaque segment peuvent être concatenées pour générer un grand nombre de variantes de chaque mot. Il faudra, selon une technique à déterminer (classification automatique), choisir un nombre convenable de références pour chaque mot. L'algorithme de reconnaissance sera alors identique à celui de l'approche globale. Dans [1], Rosenberg décrit une version monolocuteur d'un tel système qui donne de bons résultats.

*Cette étude a été financée en partie par un contrat ESPRIT.

b) Une autre approche pourrait utiliser un algorithme de reconnaissance de mots enchaînés, en remplaçant les mots par des segments, et en utilisant le lexique en guise de syntaxe. Des exemples de reconnaissance de mots enchaînés sont donnés dans [2, 3, 4]. Seul [3] aborde le problème de références multiples pour chaque mot.

2. Avec segmentation

Le mot inconnu est segmenté avant la reconnaissance. La segmentation devant être fiable, nous avons choisi de préférence une détection de noyaux vocaliques (segmentation en syllabes). Chaque segment peut ensuite être identifié soit globalement (références globales ou obtenues en concatenant des segments plus petits) soit analytiquement (chercher la meilleure séquence de sous-segments correspondant à un segment valide).

3. Apprentissage

Les segments de base seront des demi-syllabes ainsi que quelques phonèmes en position finale. Ces sous-segments doivent être choisis avec l'idée de pouvoir reconstituer n'importe quel mot du vocabulaire. Un programme interactif a permis de commencer la création d'un premier corpus (20 locuteurs, 5 élocutions par locuteur, 100 segments).

Ce premier corpus doit nous permettre d'évaluer les différentes approches décrites ci-dessus. Les performances sur un vocabulaire de 35 mots seront comparées à celles obtenues par l'approche globale (des résultats seront disponibles pour les journées d'étude). Ce corpus sera progressivement complété (100 locuteurs, un millier de segments).

[1] E. Rosenberg, L.A. Rabiner, J.G. Wilpon et D. Kahn, "Demisyllable-Based Isolated Word Recognition System", IEEE Trans. on Acoust. Speech and Sign. Proc. vol ASSP-31, n°3, juin 1983.

[2] L. Sauter, "RAPACE, un système de reconnaissance analytique de la parole continue", 4ème congrès Reconnaissance des Formes et Intelligence Artificielle, Paris, janvier 1984.

[3] C.S. Myers et S.E. Levinson, "Speaker Independant Connected Word Recognition Using a Syntax-Directed Dynamic Programming Procedure", IEEE Trans. on Acoust. Speech and Sign. Proc., vol ASSP-30 n°4, août 1982.

[4] J.S. Bridle, M.D. Brown et R.M. Chamberlain, "An Algorithm for Connected Word Recognition", JSRU Research Report n°1010, oct. 1981.

SERAPHINE : SYSTEME DE RECONNAISSANCE DE COURTES PHRASES

C. GAGNOULET, D. JOUVET : CNET LAA/TSS - LANNION - France

SERAPHINE est le nom d'un système de reconnaissance de parole, destiné à interprêter des phrases simples (grammaires régulières), construites avec des vocabulaires limités (une centaine de mots environ).

APPRENTISSAGE

Le système est dépendant du locuteur et nécessite donc une phase d'apprentissage. L'apprentissage comporte deux étapes : dans un premier temps, l'utilisateur doit décrire son application en termes de vocabulaire et de syntaxe. Il doit introduire sous forme orthographique tous les mots utilisés, ainsi que les règles de la grammaire sous une forme standard (proche de BNF). Ensuite, il doit prononcer une fois chacun des mots de l'application, ceci sous la forme de mots isolés.

PRETRAITEMENT

La séparation entre bruit et parole est effectuée par un automate travaillant sur l'amplitude du signal mesurée sur des fenêtres de 10 ms. Six coefficients cepstraux sont calculés toutes les 20 ms, à partir de spectres numériques obtenus par FFT. Chaque coefficient est codé sur 8 eb., ce qui conduit à un débit acoustique de 2400 bit/s.

L'ensemble de l'analyse acoustique a été implanté sur des processeurs de traitement de signal NEC 7720, et permet d'atteindre un traitement en temps réel.

RECONNAISSANCE

L'algorithme de reconnaissance est de type "global". Une distance acoustique est calculée entre chaque fragment de la phrase inconnue, et une concaténation de mots isolés provenant de

l'apprentissage. L'emploi du principe de programmation dynamique, et l'intégration, sous forme de réseau des contraintes syntaxiques et acoustiques, permet d'assurer en fin de phrase l'optimalité, au sens de la distance acoustique, de la séquence de mots reconnue.

Diverses heuristiques permettent de réduire l'espace de recherche des solutions, d'accepter des pauses en milieu de phrase, et de tolérer des déformations acoustiques aux frontières entre les mots dans la phrase (phénomènes de coarticulation).

L'algorithme utilisé permet un traitement séquentiel de la phrase, chaque trame de coefficients cepstraux étant traitée dès son arrivée. L'algorithme, implanté sur micro-processeur permet d'atteindre un fonctionnement proche du temps réel pour des vocabulaires réduits (facteur de branchement syntaxique de 10).

EVALUATION

Le système SERAPHINE, industrialisé sous forme monocarte par la société XCOM, a été testé sur quelques applications classiques : séquences de chiffres, numéros de téléphone, machine de bureau... L'emploi d'une syntaxe rend la procédure d'évaluation objective encore plus délicate que pour les mots isolés. On peut seulement relever un manque de robustesse du système vis à vis des conditions de prise de son, et une forte sensibilité au locuteur (bons ou mauvais locuteurs !!).

Les efforts actuels visent à rendre le système le plus indépendant possible du locuteur.

APPRENTISSAGE DES REGLES D'INTERPRETATION D'EVENEMENTS PSEUDO-PHONETIQUES

J. GUIZOL, H. MELONI : Groupe Intelligence Artificielle
Faculté des Sciences de Luminy - Marseille

1 - JUSTIFICATION DE LA METHODE

Le système que nous utilisons et la représentation paramétrique du signal de parole que nous avons choisie sont décrits dans (Méloni 82a, 82b, 84). Nous avons constaté que pour une même chaîne phonémique prononcée, la suite des évènements pseudo-phonétiques issus de la phase d'analyse de notre système différait d'un locuteur à un autre.
A l'opposé, pour un même locuteur, une stabilité apparaissait dans les évènements fournis pour diverses pronociations dans différents contextes d'une même chaîne phonémique. Par contre, certains contextes semblaient induire une caractéristique. Les règles d'interprétation de notre système devaient donc tenir compte des diverses réalisations, ce qui allourdissait grandement cette phase sans pour autant nous dispenser de devoir procéder à un ajustement des règles pour chaque nouveau locuteur.

Nous avons donc préféré "personnaliser" ces règles d'interprétation des évènements fournis par le système, au moyen d'une passe d'apprentissage.

2 - MISE EN OEUVRE

C'est la généralisation sous deux de ses aspects qui intervient dans cette phase, utilisant une stratégie conduite par les données en profondeur d'abord (Winston 75) :
1°) Une généralisation statistique qui aura pour effet de déterminer dans l'espace défini par les paramètres des évènements un pavé pour chaque phonème.
2°) Une généralisation ayant pour but, compte tenu des correspondances "suite d'évènements/suite de phonèmes", d'inférer les règles qui permettront ensuite d'interpréter les évènements.

2.1 - Découpage syllabique

Afin de rendre plus rapide cette passe d'apprentissage, nous avons essayé de réduire le plus possible les risques d'ambiguité. Pour ce faire, nous avons choisi de segmenter chaîne phonémique fournie et liste d'évènements calculée en syllabes. Pour la première, nous avons simplement affecté un niveau 0 aux voyelles, 1 aux semi voyelles, 2 aux consonnes liquides, nasales et constrictives, 3 aux consonnes occlusives. Dans ces conditions, la chaîne phonémique se traduit par une courbe en dent de scie, chaque portion décroissante déterminant une syllabe. Pour la deuxième, nous utilisons l'énergie et le premier pic du spectre pour aboutir là aussi à une courbe où chaque portion monotone croissante caractérise une syllabe.

2.2 - Superposition syllabique et détermination des groupes

Ces deux découpages étant faits, les évènements fiables, tels ceux déterminés comme étant occlusifs ou constrictifs, sont mis en correspondance de même que les segments vocaliques bien définis, les intervalles de variation de chaque paramètre étant éventuellement réajustés à chaque affectation. Les règles relatives aux autres évènements dont la correspondance avec un phonème s'opère sans ambiguité sont ensuite réajustées.

A ce stade, le système fonctionne par élimination à l'intérieur des intervalles définis par les frontières syllabiques. D'autre part, toute affectation peut être contrôlée par l'évaluation hiérarchisée des traits phonétiques du segment étudié à partir des formes linéaires discriminantes calculées, pour le locuteur considéré, dans une étape précédente (Guizol 81).

2.3 - Inférence des règles d'interprétation

Enfin, les règles contextuelles correspondant à des suppressions ou insertions d'évènements ainsi que celles rendues nécessaires par une ambiguïté d'interprétation sont inférées en cas de première apparition. Si une règle compatible existait auparavant, elle est remplacée dans le système par la plus petite généralisation (Plotkin 70). Les règles générées à ce niveau constituent un ensemble capable de prendre en compte les caractéristiques induites par les traitements précédents, les influences contextuelles et surtout, les caractéristiques de réalisation propres au locuteur.

REFERENCES

(GUIZOL 81) Guizol J., Méloni H. "Apport de l'analyse discriminante au problème de l'adaptation au locuteur en reconnaissance automatique de la parole". Session affichée - 12ème J.E.P. Montréal (1981).

(MELONI 82a) Méloni H. "Etude et réalisation d'un système de reconnaissance automatique de la parole continue", Thèse d'Etat, Université Aix-Marseille II (Février 1982).

(MELONI 82b) Méloni H., Guizol J. "A Speech Recognition System". Congrès ICASSP 82 Volume 3 - pp. 1625-28.

(PLOTKIN 70) Plotkin G.D. "A note on inductive generalization". Machine Intelligence 5 (Edinburgh University Press) pp. 153-163 (1970).

(WINSTON 75) Winston P.H. "The psychology of computer vision". New York : McGraw-Hill. (Learning structural descriptions from examples) pp. 157-209 (1975).

EVALUATION DE SYSTEMES DE RECONNAISSANCE MOTS ISOLES-MONOLOCUTEUR, AU MOYEN D'UNE CHAINE DE TEST AUTOMATIQUE

B. DUPEYRAT, D. TUAL, J.Ch. CHAZE

Centre d'Etudes Nucléaires de Saclay
91191 GIF-SUR-YVETTE CEDEX - France

L'évaluation objective de la qualité d'un système de reconnaissance de la parole est un problème crucial qui se pose aussi bien au concepteur qu'à l'utilisateur dudit système.

Il n'est en principe pas difficile, en choisissant un vocabulaire approprié et des locuteurs entraînés, de réaliser un test qui serait censé garantir le taux de reconnaissance de 98 ou 99% annoncé par la plupart des constructeurs. Mais que deviennent ces performances lorsque le vocabulaire change? lorsque le matériel est confié à un utilisateur non averti?...

La comparaison objective entre 2 systèmes ne peut se faire qu'à partir du même matériau phonétique (ou mieux à partir de plusieurs variétés de matériaux phonétiques, chacune ayant sa spécificité, car il n'y a aucune raison pour qu'un système, performant sur un vocabulaire difficile - type paires minimales - présente aussi une bonne résistance au bruit par exemple).

En attendant que le matériau phonétique soit disponible dans des banques de données et facilement accessible, nous avons constitué le nôtre sous la forme de bandes magnétiques (analogiques) contenant des enregistrements réalisés par différents locuteurs. Chaque enregistrement est constitué de plusieurs passes d'apprentissage (4 à 8) suivies de 10 séries de reconnaissance. Chaque passe ou chaque série contient tout le vocabulaire (65 ou 100 mots suivant le cas), dans un ordre "naturel" à l'apprentissage, dans un ordre aléatoire à la reconnaissance (pour éviter un phénomène d'accoutumance à la répétition des mêmes séquences). Le vocabulaire choisi est de moyenne difficulté. Le nombre de mots à reconnaître (650 ou 1000) est suffisamment grand pour que le résultat soit statistiquement significatif.

Le programme de test s'exécute ensuite sur un système de développement INTEL MDS 800; il gère le dialogue avec le système de reconnaissance et les commandes de mise en route et d'arrêt du magnétophone. Il affiche les mots reconnus en regard de ceux qui auraient dû l'être, calcule le pourcentage de reconnaissance et son évolution au cours du déroulement du test et imprime une table de confusion; il peut fournir un certain nombre d'informations supplémentaires lorsque la carte de reconnaissance le permet : score du mot trouvé, score du 2ème,... .

Le programme de test est conçu de façon que son adaptation à des systèmes de reconnaissance divers demande un minimum de modifications sur l'interface - matérielle et logicielle - entre le système de test et les cartes de reconnaissance.

Cette méthodologie de déroulement entièrement automatique du test présente des avantages notables :

- objectivité : il n'est pas possible de rectifier un mot ou de modifier (volontairement ou non) la façon de prononcer pour améliorer les résultats ainsi qu'on serait tenté de le faire dans le cas d'un test manuel.

- répétitivité et reproductibilité : ces caractéristiques sont particulièrement intéressantes en phase de mise au point pour chiffrer précisément l'amélioration apportée par la modification d'un algorithme. Il est ainsi possible de mesurer l'effet introduit par une atténuation (ligne téléphonique par exemple) ou des perturbations extérieures (bruits). Et bien sûr c'est la reproductibilité qui garantit la validité de la comparaison entre plusieurs systèmes.

Nous présenterons les résultats obtenus sur notre système (dénommé système n°1) et sur un autre système, d'origine américaine (dénommé système n°2); si l'on ne s'en tient qu'aux chiffres, ces résultats peuvent être brutalement résumés sur le tableau qui suit :

	Système n° 1 (CEA/DEIN)		Système n°2	
Nombre de mots du vocabulaire	65	100	65	100
1 passe d'apprentissage	90,5	89,5	59	39
4 passes "	94,1	94,6	71,5	66
8 passes "	92,7	(*)	75	(*)

(*) : il n'y a que 4 passes d'apprentissage sur la bande de 100 mots.

Nous espérons tester au moins une autre carte dans les semaines à venir et être en mesure d'en publier les résultats.

PROSEIDON : DETECTION AUTOMATIQUE DES INDICES PROSODIQUES CONTENUS DANS LA PAROLE CONTINUE

J. VAISSIERE, M. GILLOUX, C. TARRIDEC, M.A. SIMON
C.N.E.T. - Route de Trégastel - 22301 LANNION - France.

PROSEIDON est un programme destiné à extraire automatiquement de la phrase parlée les frontières acoustico-prosodiques présentes dans le signal. Son but est de faciliter la reconnaissance automatique de la parole continue en la divisant en unités de sens restreintes (appelées également groupes rythmiques ou mots prosodiques). Proséidon est implanté sur le système expert SERAC, en cours de développement au CNET et qui applique des techniques d'Intelligence Artificielle au problème de la reconnaissance automatique de la Parole.

Le cinquième des régles, celles qui étaient les plus aisées à formaliser, ont été intégrées et testées sur 32 phrases en Mars 1984. Ces phrases sont extraites d'un corpus de phrases dites phonétiquement équilibrées, prononcées par 3 hommes et 3 femmes pour des tests de qualité sur des vocodeurs (4 à 6 phrases par locuteur). Les 32 phrases contiennent 120 mots lexicaux (ex: "la voiture s'est arretée au feu rouge"= 4 mots lexicaux), avec une moyenne de 3,75 mots lexicaux par phrase. Les marques de segmentation de la phrase en "phonèmes", indispensables au programme prosodique pour calculer la durée des segments en séquence et pour répérer sur les zones vocaliques les valeurs du fondamental pertinentes, sont fournies par le module acoustico-phonétique mis au point au CNET, par G. Mercier.

Les premiers tests ont permis d'obtenir les résultats suivants:

1) 31 phrases (sur 32) ont été reconnues comme déclaratives et une comme interrogative (pas d'erreur).

2) les 32 phrases ont été divisées en 89 groupes rythmiques. Si la dernière syllabe d'un groupe rythmique est interpretée comme étant la dernière syllabe d'un mot lexical (e muet non compris), alors il y a deux erreurs d'interpretation (moins de 3% d'erreurs). Chaque groupe rythmique comporte en moyenne 1.3 mots lexicaux (92% des groupes comportent un ou deux mots lexicaux précédés ou non d'un certain nombre de mots grammaticaux: 64 groupes avec un mot lexical, 18 avec 2 mots lexicaux, 5 avec 3 mots lexicaux, et 2 avec 4 mots lexicaux).

3) 26 syllabes des 31 phrases reconnues comme déclaratives ont été retenues comme les syllables les plus prosodiquement dominantes de la phrase. Si on interprète la syllabe la plus dominante d'une phrase reconnue comme déclarative comme la marque de la frontière majeure il n'y a pas d'erreur. Sur les 5 phrases déclaratives pour lesquelles la frontière majeure n'a pas été détectée, 3 appartiennent au meme locuteur et 1 était insistée.

4) 31 syllabes ont été reconnues au contraire, comme prosodiquement très dominée. Si on considère qu'une syllabe prosodiquement très dominée ne peut correspondre qu'à un mot grammatical ou un e muet (21 cas), la syllabe initiale d'un dissyllabique avec un pattern prosodique P1 ou P2 (7 cas), ou la syllabe intermédiaire d'un mot de plus de 2 syllabes (2 cas), alors le programme de conversion aurait commis une erreur (sur l'adjectif postposé dans "vaisselle propre").

5) 3 syllabes ont été reconnues comme les syllabes initiales d'un pattern prosodique. Si ces syllabes sont interpretées comme des syllabes initiales de mots lexicaux, alors il n'y a pas d'erreur (il s'agit en fait des syllabes initiales de trois trisyllabiques).

Le programme donc fournit une moyenne de 2,8 informations non redondantes prosodiques par phrase (26 syllabes proposées à la fois comme frontière droite d'un mot lexical, frontière droite d'un groupe rythmique, et frontière majeure de la phrase, 31 syllabes proposées à la fois comme frontière droite d'un mot lexical et frontière droite d'un groupe rythmique, 3 syllabes proposées comme frontière gauche d'un mot lexical, 31 syllabes prosodiquement très dominées) avec un taux d'erreurs de 3 %.

MESURE DE LA RESOLUTION TEMPORELLE DU SYSTEME AUDITIF: COMPARAISON ENTRE LA DETECTION DE LACUNE ET LA FUSION DE CLICS

Ch. CAVE : LA 261 CNRS Parole et Langage, Institut de Phonétique
Université de Provence, 13621 AIX-en-PROVENCE - France.

La résolution temporelle est l'aptitude du système auditif à séparer perceptivement des évènements sonores successifs. Cette aptitude joue un rôle fondamental dans l'analyse de la structure temporelle des stimuli acoustiques. Des travaux récents ont montré l'intérêt de la mesure de la résolution temporelle pour la compréhension du fonctionnement normal et pathologique du système auditif. TRINDER (1979) montre que la m esure de la résolution temporelle peut être utilisée pour le diagnostic différentiel des lésions cochléaires et rétrocochléaires. TYLER et coll. (1982) établit une relation entre résolution temporelle et score d'intelligibilité du langage.

La technique la plus employée pour estimer la résolution temporelle est la détection de trou (gap detection) qui consiste à mesurer le plus petit intervalle de silence (dt) détectable dans un signal de durée relativement longue (PLOMB 1964, WILLIAMS et PERROT 1972).

Une autre technique consiste à mesurer le seuil de fusion de clics (HIRSH 1959) c'est à dire l'intervalle à partir duquel deux clics successifs sont perçus comme un seul clic.

Ces deux techniques fournissent une estimation de dt de l'ordre de 1 à 3 ms, comme valeur de référence pour les sujets normaux.

Toutefois il n'existe pas d'études comparant, sur les mêmes sujets, les résultats des deux techniques de mesure; ce qui permettrait de savoir si on peut utiliser indifférement l'une ou l'autre pour estimer la résolution temporelle.

Nous avons donc mesuré sur un groupe de sujets audiométriquements normaux, la résolution temporelle par la technique de la détection de trou et par la technique de la fusion de clics. Les résultats obtenus par l'une et l'autre technique sont présentés pour l'oreille droite et l'oreille gauche. On discute la corrélation entre les deux types de mesure et leur intérêt comme épreuve audiométrique supraliminaire.

HIRSH, I.J. 1959 : Auditory perception of temporal order. J. Acoust. Soc. Am., 31, 759-767.

PLOMP, R. 1964 : Rate of decay of auditory sensation. J. Acoust. Soc. Am., 36, 277-282.

TRINDER, E. 1979 : Auditory fusion : a critical interval test with implications in differential diagnosis. Brit. J. Audiol., 13, 143-147.

TYLER, R.S.; SUMMERFIELD, Q.; WOOD, E.J.; FERNANDES, M.A. 1982 : Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. J. Acoust. Soc. Am., 72, 740-752.

WILLIAMS, K.N.; PERROT, D.R. 1972 : Temporal resolution of tonal pulses. J. Acoust. Soc. Am., 51, 644-647.

UN MODELE DE NEURONE APPLIQUE AU SYSTEME AUDITIF

J.M. DOLMAZON et B. DONHOUEDE

Laboratoire de la Communication Parlée (L.A. CNRS n°368)
E.N.S.E.R.G. - 23, avenue des Martyrs - 38031 GRENOBLE cedex

Le travail que nous présentons s'inscrit dans le cadre des recherches que notre laboratoire poursuit sur l'étude du fonctionnement des différents étages du système auditif périphérique. Nous présentons ici un nouveau modèle de l'étage de génération des influx nerveux qui tient compte des plus récentes découvertes dans ce domaine. On sait, en effet, que cette étape du codage de l'information acoustique possède des propriétés fondamentales dont il faut tenir compte dans l'étude du comportement de l'ensemble.

Notre objectif a été de réaliser un modèle de neurone fonctionnant en temps réel, capable de reproduire, d'une part le comportement des signaux observés dans les fibres du nerf auditif en fonction de l'intensité de la stimulation d'entrée (activité spontanée, seuil, saturation), d'autre part certaines caractéristiques temporelles des décharges nerveuses (synchronisation des influx sur le signal incident, phénomènes d'adaptation et de recouvrement). Pour cela nous avons utilisé un dispositif à micro-processeur qui réalise la simulation sous forme programmée. Le signal d'entrée (stimulation analogique) est échantillonné par le système à micro-processeur avant d'être utilisé pour la modélisation et les potentiels d'action sont générés sous forme d'impulsions rectangulaires qui ne nécessitent pas d'étage de conversion analogique.

La structure du modèle que nous proposons appartient à la catégorie des modèles à réservoirs. Nous disposons d'un premier réservoir qui simule, dans le neurone, l'accummulation du potentiel générateur (trans-membranaire) qui au-delà d'un certain seuil de fonctionnement entraîne la génération d'un potentiel d'action. Ce réservoir est alimenté par une image de la stimulation d'entrée (vibration de la membrane basilaire). Un deuxième réservoir permet de simuler le potentiel d'adaptation qui pour chaque émission d'un potentiel d'action réduit le potentiel générateur. Nous modélisons en plus l'évolution du seuil d'excitabilité du neurone aprés l'émission de chaque potentiel d'action. Par ce moyen nous introduisons la période réfractaire du neurone qui conduit à une fréquence maximale d'émission des potentiels d'action.

Le modèle que nous avons mis au point fonctionne en temps réel et les informations qu'il délivre sont analysées par un calculateur qui permet de tracer les histogrammes d'écarts ou d'intervalles de temps et d'extraire les caractéristiques temporelles des trains d'impulsions obtenus (taux de synchronisation par exemple).

Nous présentons les premiers résultats obtenus à l'aide de ce modèle. En l'absence de stimulation d'entrée, le modèle possède une activité spontanée dont les caractéristiques statistiques correspondent bien aux résultats expérimentaux disponibles. En présence d'une stimulation tonale, le modèle reproduit, de façon satifaisante, les principaux phénomènes non linéaires : adaptation et recouvrement en début et en fin de stimulation, saturation de la réponse en fonction du niveau et de la fréquence de stimulation, et synchronisation des décharges sur le signal de synchronisation. Nous avons caractérisé ce dernier phénomène par le "taux de synchronisation" évalué par différents auteurs et nous montrons que son évolution, en fonction du temps ou de l'intensité de la stimulation, est conforme aux résultats expérimentaux obtenus dans les fibres du nerf auditif.

L'étude des différents régimes de fonctionnement nous a permis de déterminer les valeurs optimales à affecter aux différents paramètres du modèle pour pouvoir disposer d'un modèle à coefficients constants. Ce modèle sera couplé avec le modèle de membrane basilaire dont nous disposons au laboratoire.

ESTIMATION PAR LE TEST DU SEUIL DE PULSATION, DE REPRESENTATIONS INTERNES DE VOYELLES DE SYNTHESE: ETUDE DE LA DETECTION DE FORMANTS

P. ESCUDIER, J.L. SCHWARTZ

Institut de la Communication Parlée (L.A. C.N.R.S. N°368)
E.N.S.E.R.G. - 23, rue des Martyrs - 38031 GRENOBLE Cedex

Notre groupe travaille depuis plusieurs années, en collaboration avec l'Institut Pavlov de Leningrad, sur les mécanismes de traitement du spectre de voyelles synthétiques stationnaires. Nous nous sommes notamment penchés sur la mesure de seuils de détection d'"irrégularités spectrales", et nous avons montré que de très faibles pics spectraux étaient détectés, et que le seuil de détection dépendait de la forme spectrale de la référence sur laquelle naissent ces pics (1, 2, 3). Ces résultats ont conduit le groupe de l'Institut Pavlov à suggérer l'existence d'un processus de type dérivation fréquentielle du spectre d'excitation (3, 4) dont seraient responsables les mécanismes d'inhibition latérale.

Notre objectif est de réanalyser les résultats précédents à l'aide d'un outil d'évaluation des représentations internes de spectres de signaux statiques, le test du seuil de pulsation (5, 6). Il est généralement admis que cette méthode permet d'estimer le résultat de l'analyse spectrale périphérique du signal, en tenant compte notamment des effets de l'inhibition latérale (5). Un certain nombre d'études sur des signaux stationnaires complexes ont déjà été entreprises à l'aide de cet outil. Nous étudions ici essentiellement les modifications de la représentation interne d'une "pseudo-voyelle" à un formant lorsque l'on fait naître sur la pente descendante de ce formant un pic local, simulant l'apparition d'un "second formant", en mettant l'accent sur la manifestation prise dans le cadre par les phénomènes de suppression à 2 tons.

La première partie de ce travail a consisté à déterminer des conditions expérimentales favorables pour effectuer nos mesures. La méthode du seuil de pulsation, en effet, s'avère poser de gros problèmes d'utilisation pour qui souhaite faire des estimations précises. Or, pour étudier l'influence de faibles irrégularités spectrales, nous avons cherché à mesurer des différences significatives de l'ordre du décibel. Pour assurer une précision suffisante à nos mesures, nous avons donc, dans un premier temps, consacré nos efforts à l'estimation de la précision de cet outil et de la reproductibilité des mesures à travers les tests, les individus, les procédures. Nous décrivons les résultats

de cette étude préliminaire, et les solutions adoptées.

Nous montrerons ensuite comment se traduit, au niveau des représentations internes, l'émergence d'une irrégularité spectrale sur un signal référence à un formant, en fonction de l'intensité de la référence, de sa forme, de l'intensité du second pic émergeant, de sa fréquence. Les premiers résultats obtenus ne semblent pas confirmer l'hypothèse d'un processus de type dérivation spectrale : les phénomènes d'inhibition latérale, sur des pics de faible amplitude tels que ceux dont nous étudions l'émergence, paraissent de peu d'effet, et ne semblent pas à même réaliser un renforcement sensible des contrastes par rapport à l'émergence du pic secondaire dans le signal acoustique d'entrée.

BIBLIOGRAPHIE

1. LUBLINSKAYA V.V., ESCUDIER P., CARRE R. (1980) Study of the formant detection thresholds. J. Acoust. Soc. Am., 67, S1, S102.

2.LUBLINSKAYA V.V., CARRE R., ESCUDIER P. (1981) Study of some conditions determining the auditory perception of the phonetically meaning spectral cues of steady-state synthetic vowels. Actes du 1er Symposium Franco-Soviétique, p. 19-31, GRENOBLE.

3.CHISTOVICH L.A., SHEIKIN R.L., LUBLINSKAYA V.V. (1979) "Centers of gravity" and spectral peaks as the determinants of vowel quality. In "Frontiers of speech communication research" B. LINDBLOM et S. OHMAN, ed. Academic Press.

4.CHISTOVICH L.A. (1980) Auditory processing of speech. Language and Speech, Vol. 23, part. 1, p. 67-73.

5.HOUTGAST T. (1974) Lateral suppression in hearing. A psychological study on the ears' capability to preserve and enhance spectral contrasts. Doctoral dissertation. Free University, AMSTERDAM.

6.ESCUDIER P., BOULOGNE M. (1983) La méthode du seuil de pulsation en psychoacoustique : implantation et applications. Bulletin de l'Institut de Phonétique de Grenoble, Vol. 12, p. 141-153.

TRAITS ACOUSTIQUES - AIDES A LA LECTURE LABIALE

G. VILACLARA: Ecole Polytechnique Fédérale de Lausanne,
Laboratoire d'Electromagnétisme et d'Acoustique
Chemin de Bellerive, 16 - CH-1007 LAUSANNE - Suisse.

INTRODUCTION

Fournie au sens tactile par le biais d'une prothèse de substitution, l'information complémentaire donnée par certains traits acoustiques est utile à l'enfant sourd profond car ils permettent de diminuer l'ambiguïté de la lecture labiale et d'en faciliter l'apprentissage. Parmi ces traits, la fréquence fondamentale du voisement semble être un des plus importants [1].

Cette présentation concerne exclusivement la fondamentale et son extraction.

DESCRIPTION DE LA METHODE

L'extraction de la fondamentale est effectuée directement sur le signal temporel, où l'on distingue une succession de segments alternativement voisés et non-voisés. Un segment est considéré comme voisé s'il peut être décomposé en sous-segments plus petits de forme semblable. Les pseudo-périodes ainsi définies s'identifient par une relation d'anamorphose.

Après avoir subi une transformation non linéaire accentuant la fondamentale du voisement [2], le signal de parole est filtré passe-bas. On établit ensuite la liste exhaustive des paires de pseudo-périodes hypothétiques. Chacune des pseudo-périodes est définie par une combinaison de sections temporelles successives, bornées de part et d'autre par un passage par zéro à pente positive. Chaque pseudo-période d'une paire est ensuite normalisée en durée par subdivision en 16 sous-segments. Comme il est fort peu probable que le nombre des échantillons constituant la pseudo-période soit un multiple entier de 16, on définit deux types de sous-segments de longueurs distinctes (la longueur d'un sous-segment est donnée par le nombre d'échantillons qui le constituent). En répartissant de façon uniforme et dans des proportions adéquates les grands sous-segments parmi les petits, on obtient la forme globale de la pseudo-période considérée. Les 16 éléments de comparaison sont donnés par la valeur moyenne des échantillons englobés par chaque sous-segment. On considère que l'hypothèse qui satisfait au mieux les critères de ressemblance contient deux périodes fondamentales.

En imposant que le nombre d'échantillons contenus par un sous-segment quelconque soit une puissance de deux; il est possible d'effectuer une sous-segmentation préalable de la partie du signal de parole analysée. En effet, si l'on considère le point initial au début de la recherche, une périodicité liée à un petit nombre d'échantillons nécessite des sous-segments de courte durée. Plus la période est longue plus la sous-segmentation porte sur un grand nombre d'échantillons et plus elle s'éloigne du point initial. Un code constitué d'une succession de sous-segments croissants de façon monotone, permet de réduire la quantité d'information nécessaire à l'extraction, ce qui constitue un avantage non négligeable (la capacité mémoire de processeurs rapides spécialisés pour le traitement numérique est actuellement assez faible).

CONCLUSION

L'algorithme fonctionne en simulation sur un miniordinateur et fournit des résultats satisfaisants. Le codage effectué permettra son implantation sur un système à microprocesseurs.

PUBLICATIONS

[1] M. Rossi, D. Schneuwly, "Traitement des informations acoustiques en vue d'une substitution tactile", Bulletin de l'Audiophonologie, no 6-7, vol. 16, 1983, pp. 801-814.

[2] Wolfgang Hess, "Pitch Determination of Speech Signals", Springer Verlag, Berlin Heidelberg New York Tokyo, 1983, pp. 166-173.

BANDE PASSANTE, PAROLE ET LANGUE ETRANGERE

J. JANDOT : D.R.L. de Paris VII

Il est fréquent de lire , chez de multiples auteurs , qu'une bande passante limitée suffit pour la transmission de la parole: le téléphone en fournit un exemple quotidien.

Cependant , la complexité de certains messages vocaux - pour le récepteur , du moins - rend un canal limité à 300-3000 Hz (en simplifiant) , et affecté de distorsions et bruits divers , nettement insuffisant pour une transmission sûre. La redondance ne joue pas forcément: un nom propre inconnu est souvent mal entendu ; on doit le répéter , l'épeler au besoin. Les personnels des télécommuni--cations utilisent un code auxiliaire ("A" de Anatole , "B" de Berthe , etc...) pour éviter les erreurs éventuelles. La procédure aérienne donne lieu au même type de codage secondaire.

Si l'on considère par ailleurs qu'une évaluation quantitative (en pourcentage) ne reflète pas la réalité de la communication parlée , car la non-perception d'un seul mot peut rendre une phrase entière inintelligible , on est obligé de constater que la situation ordinaire de communication est sans doute idéalisée dans les analyses linguistiques classiques : la connaissance des phonèmes d'une langue , et même du système phonétique et phonémique , n'assure pas (avec ou sans redondance) une communication sûre dès lors que le canal est limité à la zone conversationnelle de type téléphonique.

Une expérimentation de près de dix ans auprès d'enfants et d'adultes apprenant l'anglais(techniques audio-orales et audio-visuelles) a permis de mettre en évidence le rôle essentiel des fréquences , souvent considérées comme inutiles , comprises entre 6000 et 12500 Hz .

L'utilisation d'enceintes acoustiques spécialement étudiées pour un local de classe donné , et un système de filtres (trois voies commutables et dosables séparément) a confirmé l'hypothèse selon laquelle la présence de fréquences aiguës (dans les limites indiquées ci-dessus) permet une meilleure intelligibilité de la langue étrangère , et une réduction spectaculaire de certaines confusions , notamment entre sons consonantiques voisins. (Ceci est valable pour l'anglais britannique , moins pour l'américain - trop souvent confondus dans les études qui se veulent générales...).

Des Anglais ont également été soumis à des tests: les résultats sont identiques dès qu'on ajoute au signal utile un bruit de fond parasite , même de faible intensité. Par contre , la diffusion des fréquences graves (inférieures à 200 Hz) doit rester à un niveau assez faible (sauf au casque , s'il est doté d'une bande passante assez large). Ces résultats concordent avec les chiffrages avancés par BARTH et coll. (XI emes JEP / Strasbourg. 1980. GALF. p 71)pour des malentendants.

Ces expériences conduisent à reconsidérer le problème des normes pour appareils audio d'enseignement des langues (et de leur application). Elles posent également la question des critères retenus pour la transmission sûre des messages vocaux.

Elements de bibliographie:

JANDOT.J. " Audition , apprentissage du langage et enseignement" , in. "Journée d'études: parole et déficiences auditives" .Bordeaux. GALF (GSO) .1983.

JANDOT.J. " Moyens électro-acoustiques et enseignement de l'anglais". Thèse de Doctorat d'Etat. Département des Recherches Linguistiques de PARIS VII. 1983.

PROPOSITION DE MOTS DANS UN SYSTEME DE RECONNAISSANCE DE MOTS ISOLES MULTILOCUTEURS : UNE APPROCHE EN VUE DU TRAITEMENT DES GRANDS VOCABULAIRES

Ph. LOCKWOOD: Laboratoires de Marcoussis - C.R. C.G.E.
Route de Nozay - 91460 MARCOUSSIS - France.

Les résultats de Shipman et Zue [1] montrent que la segmentation d'un mot en différentes classes phonétiques grossières réduit le nombre de candidats mots dans un système de reconnaissance de mots isolés dépendant du locuteur. Le but de ce travail est d'inclure dans un système de reconnaissance de mots isolés multilocuteur [2] un algorithme de préclassification de mots. Chaque mot est segmenté en 3 classes phonétiques : voyelle, fricative non voisée, plosive non voisée. Le but est d'obtenir le moins de décompositions possibles pour un même mot afin de réduire les risques de mauvaise préclassification et d'avoir un détecteur de voyelles fiable pour un travail ultérieur.

Le contour d'énergie du mot (fréquence d'échantillonnage : 8000 Hz) est utilisé pour la détection début-fin de mots. Ce contour d'énergie du mot ainsi qu'une mesure de stabilité spectrale sont combinés pour proposer des candidats voyelles.

Un algorithme de reconnaissance des formes est utilisé pour effectuer, entre les marqueurs de début et fin de mot, une détection silence-voisé-non voisé indépendante du locuteur [3] ; un label est attribué toutes les 8 millisecondes (fenêtre de 16 ms). Un automate est utilisé sur les labels ainsi obtenus pour segmenter le mot en 3 classes : voisé, fricative non voisée et candidat plosive non voisée. Les plosives non voisées sont retenues après examen (toutes les 4 ms, fenêtre 16 ms) du spectre du signal suivant le silence.

Le résultat de cette segmentation est ensuite combiné avec les candidats voyelles; le mot est ainsi segmenté en 3 classes phonétiques. Les insertions de consonnes sont traitées (il s'agit principalement de nasales, liquides, fricatives voisées fortes, plosives voisées fortes). Un sous ensemble de paramètres "invariants" [4] a été automatiquement déterminé par une procédure en pas à pas ; ces paramètres ont été choisis pour obtenir un groupement optimal des voyelles et des consonnes.

Une phase d'apprentissage est nécessaire afin de prendre en compte les variabilités et les particularités de prononciation des locuteurs. 20 locuteurs et 3 élocutions par locuteur ont été utilisés. 41 décompositions différentes sont obtenues pour le mot

vocabulaire de 35 mots [2]. Lors de la phase de reconnaissance, un mot inconnu est segmenté et tous les mots qui ont eu la même décomposition sont proposés.

Ce système a été testé pour 13 locuteurs n'ayant pas participé à l'apprentissage :5 mots en moyenne sont proposés. Il y a eu une erreur pour mauvaise préclassification. Le détecteur de voyelles a 2.2 % d'insertions et O.2 % d'élisions. Les performances globales du système [2] n'ont pas été modifiées.

[1] D.W. Shipman, V. Zue, "Properties of large lexicons : implications for advanced isolated word recognition systems" 1982 IEEE ICASSP Conference.

[2] B. Flocon, N. Briant, "SYRIL : système temps réel de reconnaissance de mots indépendant du locuteur" 4ème Congrès AFCET, Reconnaissance des Formes et Intelligence Artificielle 1984.

[3] L. Siegel, "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier". IEEE Trans. vol ASSP-27 FEB 1979.

[4] R.A. Cole, R.M. Stern, M.S. Philips, S.M. Brill, A.P. Pilant, P. Specker, "Feature based speaker independent recognition of isolated english letters" 1983 IEEE ICASSP Conference.

SYSTEME EXPERT DE DECODAGE ACOUSTICO-PHONETIQUE ET INVARIANCE

N. CARBONELL, D. FOHR, J.P. HATON, J.M. PIERREL
CRIN - Université Nancy 1 - BP 239 - 54500 VANDOEUVRE

F. LONCHAMP
Institut de Phonétique - Université Nancy 2 - 54000 NANCY

Dans le cadre de l'élaboration d'un système-expert de décodage acoustico-phonétique, nous avons été amenés à nous intéresser à l'invariance. L'expertise choisie est la lecture de spectrogrammes.

Le protocole d'acquisition de l'expertise est le suivant :

- l'expert décode oralement un spectrogramme puis commente ses erreurs (il ignore l'énoncé représenté) ; analyse et commentaires sont enregistrés,
- ensuite, les bandes sont analysées en dehors de sa présence, afin de dégager les indices et les règles d'identification qu'il utilise, ainsi que ses stratégies,
- enfin, les conclusions de cette étude sont soumises à son examen critique.

Pour vérifier leur pertinence, on implémente les règles, au fur et à mesure de leur découverte, dans un moteur d'inférences écrit en LISP et on les teste ; à noter que l'acquisition des indices est faite de manière conversationnelle, pour l'instant, par un non expert, à partir du spectrogramme.

Dans un premier temps, l'expert a décodé une trentaine de spectrogrammes représentant des énoncés lus par un locuteur expérimenté. Son taux de reconnaissance des consonnes avoisine 90%. Les règles d'identification qu'il utilise sont largement contextuelles tant en ce qui concerne l'émission que la validation d'hypothèses phonétiques. En effet, l'expert ne propose aucune interprétation, pour un segment donné, sans avoir, au préalable, déterminé la classe phonétique des segments contigus ; par exemple, si l'un des contextes du segment qu'il étudie est vocalique, il caractérise ce contexte à l'aide de critères articulatoires.

Dans un deuxième temps, l'expert a analysé 50 spectrogrammes représentant des phrases toutes différentes empruntées au corpus de Combescure [Combescure, 1981] et prononcées de manière naturelle* par 5 locuteurs (non entraînés) différents, à raison de dix phrases par locuteur. L'expert a segmenté et identifié correctement 80% des consonnes. Ont été considérés comme correctement identifiés les segments pour lesquels il a proposé deux interprétations phonétiques, dont la bonne.

L'analyse de la démarche de l'expert a montré que la plupart des règles mises en évidence au cours de la première étape du travail restaient valables mais

---

* Pour éviter une lecture des phrases, on a demandé aux locuteurs de mémoriser chaque phrase puis de la prononcer.

qu'elles devaient être complétées pour tenir compte des phénomènes de co-articulation, plus importants dans ce corpus que dans le premier : nasalisation des occlusives précédées d'une nasale, apparition de formants entre 1000 et 2000 Hz dans les segments fricatifs voisés situés entre deux voyelles, etc. Ces altérations sont dues essentiellement au fait que les énoncés n'ont pas été lus mais prononcés de manière naturelle à un rythme relativement rapide et que les locuteurs manquent d'expérience. A noter que l'adaptation de l'expert aux différents locuteurs est très rapide (une phrase suffit) et que les variations sont moins importantes d'un locuteur du second corpus à l'autre, qu'entre le locuteur entraîné du premier corpus (rythme régulier, articulation nette, intensité relativement constante) et ceux du second (articulation médiocre, variations d'énergie importantes). Signalons chez ces derniers un nombre non négligeable de prononciations incorrectes.

Actuellement, nous étudions les analyses et les commentaires de l'expert relatifs au second corpus, afin de compléter les règles acoustico-phonétiques d'identification mises en évidence au cours de la première étape : leur nombre est relativement restreint.

Parallèlement, nous avons regroupé les représentations (sur les cinquante spectrogrammes) de toutes les occurences d'un même phonème (consonne) avec les contextes droit et gauche correspondants, afin de contrôler la pertinence des règles utilisées par l'expert, de les affiner et de déterminer des critères d'identification des consonnes indépendants du locuteur, pour les différents contextes possibles.

Par ailleurs, nous implémentons des algorithmes de détection des indices qui constituent les prémices de ces règles ; nous les mettons au point et nous les testons sur une dizaine de phrases du second corpus.

En ce moment, nous travaillons sur les occlusives. Pour donner au système des informations analogues à celles sur lesquelles s'appuie l'expert, nous avons mis au point, avec sa collaboration, un algorithme de construction de spectrogrammes numériques que l'on peut éventuellement visualiser sur écran TV.

(i, a, u) ? PAS SI FOU ? OU LES LEVRES DES CONSONNES MAXIMISENT-ELLES L'ESPACE ACOUSTIQUE DES VOYELLES ?

Ch. ABRY et L.J. BOE: Institut de Phonétique, Institut de la Communication Parlée, GRENOBLE, L.A. C.N.R.S. 368
R. DESCOUT : R.C.P., CNET - LANNION

Nos travaux sur le trait d'arrondissement en français se situent, au moins implicitement depuis notre premier corpus (1978 in ABRY & al., 1980, p. 111), dans une approche qui recherche en priorité les manifestations de la structure sémiotique naturelle (BRØNDAL, 1943, chap. III) sur ses limites : soit entre la distinctivité maximale (LILJENCRANTS & LINDBLOM, 1972) et la neutralisation. Après avoir testé nos hypothèses sur la structure articulatoire (ABRY & BOË, 1981-1982; et ici même ABRY & BENOIT) nous voici aux prises avec la structure acoustique de [rond].

DONNEES

Nous avons récidivé : comme en 1978, pour la géométrie des lèvres, c'est un locuteur féminin[1] qui nous livre ces premiers résultats. L'enregistrement utilisé ici est la bande son d'un film 35 mm (100 images/sec.), réalisé pour nous au CNET par l'équipe du SIRP. Après numérisation du signal (à 16 kHz), les formants F_1, F_2, F_3 ont été mesurés (FFT et cepstre), dans la partie centrale la plus stable des voyelles (décision prise sur sonagramme numérique).

Le corpus comprend les oppositions minimales i~y et ɛ~œ, soit les extrêmes de [± rond], situés dans le cadre vocalique en leur ajoutant u et a. Les consonnes nécessaires pour étudier la coarticulation représentent les principales cibles aux lèvres : p, f, ʃ, s; parmi toutes les autres, réputées neutres labialement, nous avons choisi t et k, pour leurs effets mandibulaires (sur les lèvres) respectivement fermant et ouvrant (nous aurions pu pour cet effet-ci utiliser le R : nous ne l'avons pas retenu pour des raisons de segmentation). La phrase porteuse de nos logatomes était : "C'est # C_1 V_1 C_2 V_2 ? # ça ?", avec C_2 V_2 combinant les phonèmes choisis. C_1 est uniformément une dentale et V_1 présente le contraste dynamique maximum, au niveau labial, avec V_2 (ex. V_1 = u si V_2 = i). Chaque logatome est répété six fois.

CONTRASTES SUR LE CARRE DE L'ARRONDISSEMENT i:y::ɛ:œ

Nous avions prévu pour les oppositions minimales, le contraste maximal en

[1]. Née à Paris, le 11 Mai 1947, habite Nanterre jusqu'en 1967. Puis vient en Haute-Savoie (3 ans) et enfin à Grenoble (depuis 1970). Accent parisien peu marqué.

combinaison paradigmatique [+ haut] : avec i~y; le contraste minimum avec ɛ~æ [+ bas]. Et ceci est bien généralement le cas, tous contextes confondus (cf. fig. 1; ellipses à 90%).

Mais les effets des combinaisons syntagmatiques sont très divers. Ainsi si~ʃy représente bien le contraste maximum pour les voyelles hautes et sɛ~ʃæ est dans les moins bons (avec t et k). Les meilleures paires minimales sont, dans l'ordre : pɛ~pœ, si~sy, ʃi~ʃy, ki~ky et fɛ~fæ. La structure acoustique peut donc être remarquablement préservée pour certains contextes qui semblent à première vue articulatoirement contraires. Ainsi malgré la protrusion, dans ʃi, l'aire aux lèvres du i est suffisamment maintenue par l'éversion due à ʃ (ABRY & al., 1980, p. 155). Plus surprenant : l'aperture due à p (mais aussi à f) permet très bien aux voyelles ouvertes ɛ~œ de manifester leur opposition, et ce malgré un jeu des lèvres plus réduit à mandibule ouverte.

CONTRASTES VOCALIQUES MAXIMAUX : CONTRASTES LABIAUX ?

Ayant constaté que les quatre coins de la structure sémiotique, manifestant l'opposition minimale sur le trait rond, sont tenus par des consonnes qui ont un effet labial (s, ʃ, p ou f; cf. fig. 1), nous pouvons aussi nous demander comment cette structure, propre au français, se situe dans l'espace vocalique maximal.

Nos prévisions articulatoires pour i donnaient s optimisant : ce qui se vérifie dans la structure acoustique. Par contre, nous avions donné labialement ʃu gagnant : or, acoustiquement, fu lui est préféré, sans doute essentiellement parce que cette consonne laisse davantage de liberté à la position linguale de u que ʃ (ajouter que ce dernier ouvre légèrement le u; f par contre ne le recule guère se laissant presqu'entièrement assimiler en protrusion). Pour a, la prévision articulatoire correspond à l'effet acoustique : pa est bien (avec fa) le plus extrême en ouverture. A cette réserve près que l'on fasse confiance aux auteurs qui en font la consonne de plus grande aperture mandibulaire (revue in BOGNAR, 1982, p. 61; pour notre locuteur cf. p. 319 [p] et 337 [f]), on nous accordera que nous semblons tenir avec pa, si et fu les effets optimisants extrêmes des consonnes sur le triangle vocalique.

On sait tout le prix qu'on a pu attacher à la prédiction structurale en phonétique par optimisation de cet espace des voyelles (LILJENCRANTS & LINDBLOM, 1972), de ces voyelles en fonction des consonnes, enfin des suites CV (cf. en dernier lieu LINDBLOM & al., 1984). Sera-t-il indifférent à cette problématique de savoir que les effets consonantiques des lèvres tiennent les marches du pays des voyelles ?

REMERCIEMENTS :

A Michèle DUVERNEUIL, notre locutrice; au service SIRP des PTT; à Feng GANG

pour ses programmes de traitement du signal; et à Christian BENOIT pour son éditeur; enfin à Dominique VUILLET pour la présentation du texte.

REFERENCES

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

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

BOGNAR E. (1982),
Espaces et cibles mandibulaires ... - Thèse de 3e Cycle, Grenoble.

BRØNDAL V. (1943),
Essais de linguistique générale. - Copenhague.

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

LINDBLOM B., MAC NEILAGE P. & STUDDERT-KENNEDY M. (1984),
The biological bases of spoken language. - San Francisco.

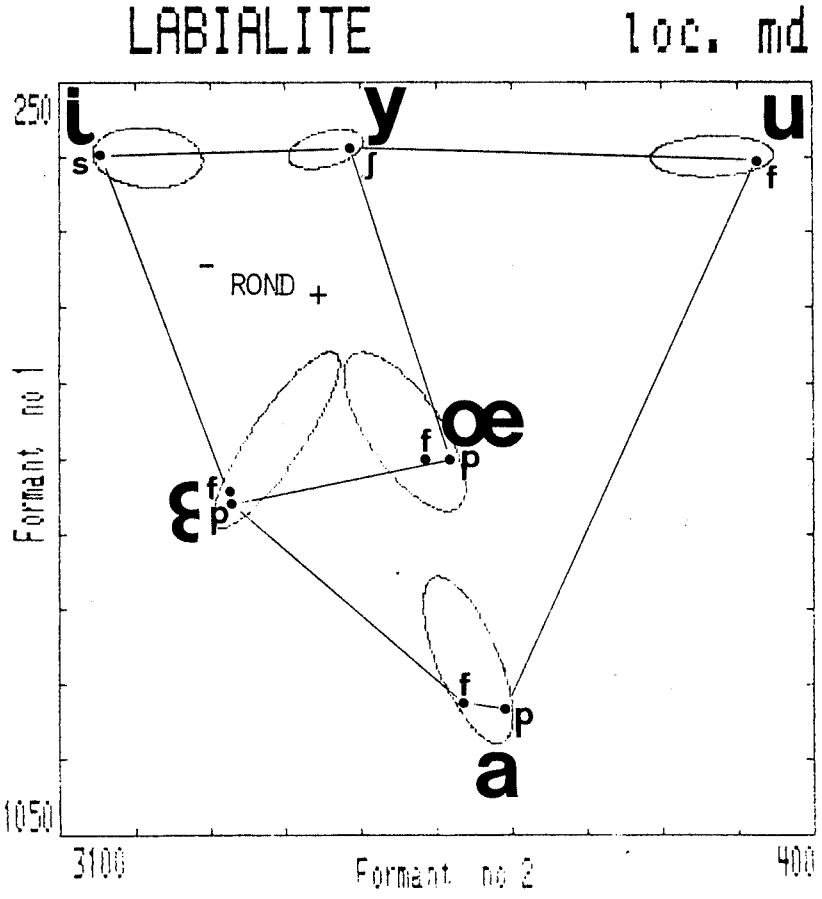


Figure 1

Ellipses de dispersion acoustique des voyelles. Les contextes consonantiques limites sont les sommets du carré de l'opposition minimale d'arrondissement et les extrêmes du triangle vocalique.

PROBLEMES D'ISOCHRONIE RECONSIDERES A LA LUMIERE DES DONNEES SUR L'ACQUISITION DU LANGAGE

G. KONOPCZYNSKI : CNRS - Laboratoires de Phonétique de BESANCON et de STRASBOURG.

Les problèmes d'isochronie en phonétique ont été à l'ordre du jour pendant de longues années, essentiellement entre 1970 et 80. (cf. bilans de Lindblom 1978, d'Oller 1979). Mais, malgré de nombreux travaux, le phénomène s'avère peu clair au niveau du langage adulte. Il a donc été suggéré, notamment au IX°I.C.P.S. (Symposium 5) de prendre le problème à son origine, en examinant ce point dans le langage enfantin.

La plupart des langues ne présentent pas d'égalité syllabique, notamment en raison de l'allongement final(AF). L'existence de celui-ci est cependant loin de faire l'unanimité des chercheurs, car si dans un nombre important de langues les voyelles (ou syllabes) finales sont allongées de 15 à 30% par rapport aux non-finales (cas de l'anglais, allemand, suédois, russe, italien, espagnol, français) il existe aussi des langues telles le finnois, l'estonien et le japonais qui n'ont pas d'AF. De ce fait, certains chercheurs ont estimé que l'AF serait inhérent aux mécanismes de production, donc contraint physiologiquement et par conséquent inné, alors que d'autres expriment l'idée que l'AF serait un comportement de production acquis. Pour tenter d'apporter une contribution à ce débat, nous nous sommes tournés vers les indications fournies par l'étude des toutes premières émissions vocales qui ne soient ni cris, ni pleurs, ni vocalisations avec structure de vocoïdes isolés, mais qui soient des structures de type CV. En effet, si aucun allongement n'est repérable dans les premiers énoncés et s'il apparaît progressivement, une explication par des causes physiologiques serait difficile à admettre et il faudrait opter pour un comportement acquis.

L'étude porte sur les rares résultats disponibles dans la littérature (uniquement pour l'anglais) et sur nos propres résultats concernant des enfants français de 8 à 24 mois. Dans les tous premiers énoncés CV (8-10 mois) il y a isochonie totale, c'est-à-dire que ni la longueur de l'énoncé, ni la place de la syllabe dans l'énoncé (finale/non fin.) n'influent sur la durée syllabique, pas plus d'ailleurs que les phénomènes de proéminence qui sont encore absents des premières émissions. La comparaison entre syllabes fin/non fin. montre un ratio généralement inférieur à 1.20, aussi bien en anglais qu'en français, donc non perceptible, puisque selon Rossi (1972) et Klatt (1976) il faut au moins 20% d'allongement pour

qu'il soit perceptible. De même, la différence de durée syllabique entre les énoncés brefs et les énoncés longs est statistiquement non significative. Dans les mois suivants, l'isochronie initiale est peu à peu perturbée : l'on montrera quel est l'ordre d'acquisition des facteurs perturbants, puisque dans les premières formes lexicales clairement identifiables, l'isochronie initiale a déjà disparu et à 2 ans et demi les rapports entre Fin. et non Fin. sont comparables à ceux des adultes et n'évoluent pratiquement plus.

KONOPCZYNSKI G. (1984) : ALLONGEMENT FINAL : CONTRAINTE PHYSIOLOGIQUE OU COMPORTEMENT ACQUIS ? QUELQUES NOTES A LA LUMIERE DES DONNEES DE L'ACQUISITION DU LANGAGE. T.I.Ph.S. 16, sous presse.

EDISIG : ENCORE UN EDITEUR DE SIGNAL ?!!

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

I - ATTAQUE DE LA BANQUE

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

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

EDISIG tente de remplir ce rôle.

II - LIMITES DES EDITEURS EXISTANTS

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

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

III - AVANTAGES D'EDISIG

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

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

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

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

IV - FICHE TECHNIQUE

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

.....

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

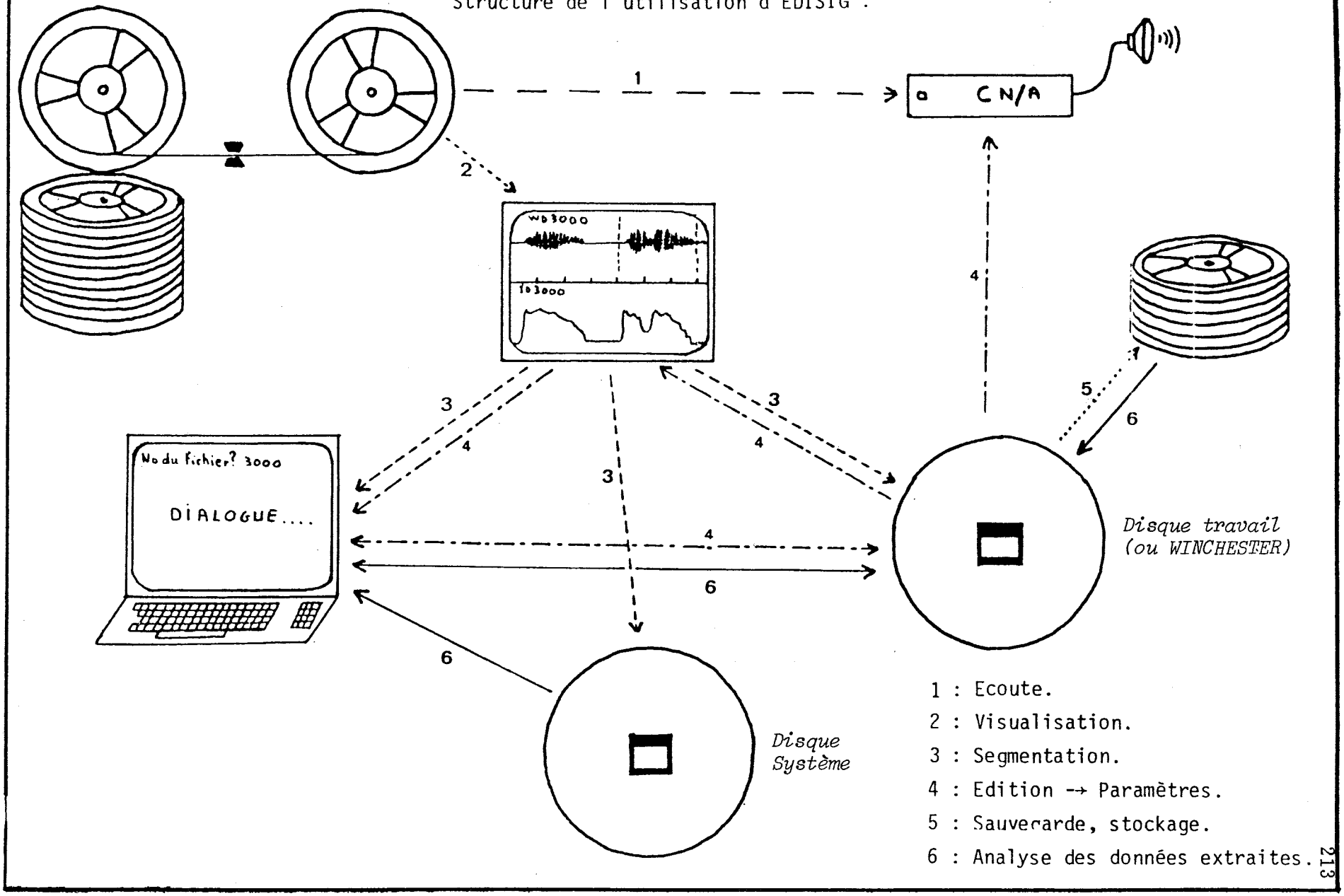
Analyse cepstrale, visualisation sonographique et détection des formants. - Séminaire GALF-GRECO Analyse du signal de parole. Paris.

---

*. Contrat CNET; Caractéristiques individuelles

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

Structure de l'utilisation d'EDISIG :



COARTICULATION ANTICIPANTE ET COARTICULATION RETENTRICE EN FRANCAIS: PHYSIOLOGIE DE QUELQUES INDICES ELECTROMYOGRAPHIQUES

J-F.P. BONNOT
Département de français de l'Université d'Amsterdam
C. CHEVRIE-MULLER, G. GREINER, C. GUIDET et B. MATON
Institut National de la Santé et de la Recherche Médicale - U3

A partir d'un corpus composé de 20 énoncés français, nous avons examiné la question de la migration des traits de labialité et de nasalité hors du domaine segmental qui leur est imparti dans les descriptions phonologiques classiques. Nous avons recueilli les potentiels électromyographiques des muscles élévateur du voile (levator veli palatini: LP) et orbiculaire des lèvres (orbicularis oris: OO), pour 5 sujets (3 femmes et 2 hommes). La méthode est exposée ailleurs en détail (Bonnot et al. 1980).
Les mesures effectuées sur les tracés ont trait aux durées d'activité et de suppression d'activité musculaires et à divers temps de latence, mettant en rapport les signaux acoustique et électromyographiques.

ANTICIPATION DE LA LABIALITE

Dans des séquences du type $-VC_6[y]-$ où C_6 est une suite consonantique non labialisée, le timing de mise en oeuvre de la protrusion (OO) est dépendant de contraintes biomécaniques. La présence de [i] (antagoniste de [y] du point de vue labial) favorise l'expansion de la labialité. Ces résultats sont en accord avec Sussman et Westbury (1981). Nous infirmons d'autre part l'hypothèse maximaliste de migration du trait: il existe une nette variabilité interpersonnelle, l'anticipation ne débutant pas systématiquement au commencement de la séquence non labialisée.

ANTICIPATION DE LA NASALITE

Dans des séquences $-Coral_4Vnas-$, c'est [ɛ̃] qui favorise la suppression la plus importante et la plus précoce de l'activité de LP. De plus, le temps de latence n'est jamais supérieur à 17% de la durée de la suite consonantique, ce qui reflète un important cadrage de l'expansion de la nasalité.
Dans des réalisations du type $[e^3_1n]$; $[a^4_3n]$; $[a^3_0\tilde{a}]$, il convient de tenir compte (a) du nombre et de l'importance des frontières (limites syllabiques et frontières de mots) et (b) du poids physio-

logique des articulations.
(Nous confrontons nos données et celles de Benguerel et al.(1977). Sur certains points, nos investigations confirment les leurs, sur d'autres, elles les infirment. C'est notamment le cas en ce qui concerne le nombre de segments touchés par l'anticipation de nasalité).

RETENTION DE LA NASALITE

Les séquences $[\tilde{a}a_0^3]$ mettent en lumière une action de la longueur de la suite vocalique. On ne peut pas parler de persistance de la nasalité au-delà du domaine segmental nasal, mais le timing est réorganisé à l'intérieur de ce cadre: l'activité électromyographique de LP reprend d'autant plus tardivement que le nombre de voyelles subséquentes est plus grand. Ceci indique vraisemblablement que la coarticulation rétentrice n'est pas uniquement mécanique, mais qu'elle bénéficie également d'une programmation d'assez haut niveau.

REFERENCES

BENGUEREL A.P., HIROSE H., SAWASHIMA M. et USHIJIMA T. "Velar
1977 coarticulation in French: a fiberscopic/electromyographic study" Journal of Phonetics 5/2:149-168.

BONNOT J-F.P., CHEVRIE-MULLER C., GREINER G. et MATON B. "A propos
1980 de l'activité électromyographique labiale et vélaire durant la production de consonnes et de voyelles nasales en français" Travaux de l'Institut de Phonétique de Strasbourg 12: 177-224.

SUSSMAN H.M. et WESTBURY J.R. "The effects of antagonistic gestures
1981 on temporal and amplitude parameters of anticipatory labial coarticulation" Journal of Speech and Hearing Research 46: 16-24.

NASALITE ET ORALITE VOCALIQUES EN FRANCAIS : ETUDE CINERADIOGRAPHIQUE, PREMIERS RESULTATS.

J.P. ZERLING : Institut de Phonétique - Université de Strasbourg
22, rue Descartes - 67084 STRASBOURG CEDEX - France.

La nasalité vocalique est un phènomène phonétique encore relativement mal connu. Les études articulatoires descriptives portant sur le français sont rares (BRICHLER, 1970, Thèse 3è cycle). Les études physiologiques commencent à porter des fruits (BONNOT et al, 1983, ICA), et les études acoustiques par modélisation articulatoire (MAEDA, 1983, ICA) et par analyse, synthèse et étude de perception (LONCHAMP, 1979, Verbum), ont jeté récemment quelque lumière sur les problèmes posés.

Notre approche articulatoire et acoustique est fondée sur la cinéradiographie. Elle s'inscrit dans le cadre d'un travail plus vaste d'observation de phénomènes de coarticulation, dont une première partie a déjà été réalisée (ZERLING, 1979, thèse 3è cycle). Dans la présente étude, nous comparons chez deux sujets diverses réalisations vocaliques nasales et orales prononcées dans les logatomes suivants : /tṼ/ (ɛ̃, œ̃, ɑ̃, ɔ̃) et dV(d) (ɛ, œ, a, ɑ, ɔ, o).

COMPARAISON DES VOYELLES NASALES ET ORALES

Deux caractéristiques articulatoires distinguent toujours très nettement, à l'intérieur de chaque couple, la nasale de l'orale :

- L'abaissement prévisible du voile du palais qui adopte une forme courbe, signe d'une relative décontraction par rapport à sa position haute.
- Une position plus reculée de la langue pour la nasale qui entraîne une constriction pharyngale importante. Contrairement à toute attente, ce recul n'est pas nécessairement accompagné d'un abaissement du dos de la langue.

Un troisième paramètre, l'angle maxillaire, semble directement lié à la nasalité (légère fermeture), mais dans une proportion trop faible pour être distinctive à coup sûr.

En ce qui concerne la position des lèvres et du dos de la langue, on note que [ɛ̃] et [œ̃] sont relativement proches de [ɛ] et [œ], excepté pour un des sujets qui neutralise l'opposition ; que [ɑ̃] est totalement labialisée, et que [ɔ̃] est réalisée avec une position linguale voisine de celle de [ɔ] et une labialisation identique à celle de [o]

.../...

COMPARAISON DES VOYELLES NASALES

Si l'on considère la forme globale de la langue et l'angle maxillaire, les nasales constituent deux paires : [ɛ̃ - œ̃] et [ɑ̃ - ɔ̃]. La première est caractérisée par une cavité buccale plus fermée. La distinction entre deux voyelles d'une paire est obtenue en modifiant le degré de labialité. Pour [ɛ̃] et [œ̃] on procède comme pour les voyelles orales correspondantes : non labiale et labiale. En revanche, pour [ɑ̃] et [ɔ̃], la distinction n'a pas d'équivalent pour les voyelles orales : les deux voyelles étant déjà labiales, [ɔ̃] doit être "surlabialisée", c'est-à-dire que la protrusion, et surtout le rapprochement des lèvres, sont supérieurs à ce qu'on pourrait attendre étant donné le degré d'aperture buccale. Ce phénomène semble être unique pour les voyelles du français et il remet en cause, au moins pour nos deux sujets, l'opposition binaire +lab/-lab qui devient ternaire : ++lab/+lab/-lab.

Ce travail doit être complété maintenant par une analyse acoustique. Nous espérons pouvoir dégager quelques différences spectrales significatives et les mettre en relation avec les observations articulatoires. Par ailleurs, nous menons actuellement une étude sur la nasalisation des voyelles orales par le contexte, afin de voir en quoi elles se distinguent des voyelles nasales.

SYNTHESE A FORMANTS DE HAUTE QUALITE : SOURCE VOCALE ELABOREE

S. AWAD et B. GUERIN : Institut de la Communication Parlée. CNRS LA N°368
ENSERG - 23, rue des Martyrs - 38031 GRENOBLE CEDEX - France

La plupart des études sur le fonctionnement de l'appareil phonatoire humain et sur la synthèse de la parole se sont plus particulièrement concentrées sur le fonctionnement et les caractéristiques du conduit vocal dans un premier temps, négligeant un peu la source vocale. Il est vrai que les informations sur la source sont trés difficiles à obtenir. Pendant longtemps, on s'est contenter d'exciter les syntétiseurs à formants par un simple signal impulsionnel mis en forme à l'aide de circuits spéciaux. Mais depuis une dizaine d'années, les chercheurs ont songé à améliorer le signal d'excitation afin d'obtenir une parole de synthèse plus naturelle.

En effet, des études ont montré que la qualité de la parole synthétique dépendait des caractéristiques temporelles de la source d'excitation, ce que nous vérifierons par la suite. Le modèle de la source d'excitation que nous avons utilisé est basé sur celui proposé par FANT (1979,1980) où le signal de débit glottique est approché par des segments de cosinus. En utilisant les résultats des caractéristiques de forme de l'onde de débit d'un modèle à 2 masses, nous avons pu établir les relations permettant de commander le modèle de FANT de l'onde de débit avec les paramètres P_s et Q du modèle à 2 masses (soit respectivement la pression subglottique et la tension des cordes vocales) (AL ANSARI 1981). Ainsi, le signal d'excitation avait-il les mêmes caractéristiques temporelles : Q.O., Q.D., F_0 que celui du modèle à 2 masses.

Sept voyelles isolées:/a e o u œ i y/ ont été synthétisées d'une part avec la source vocale élaborée, d'autre part avec une simple source d'impulsions mises en forme. Le synthétiseur utilisé est du type à formants série avec 5 formants. Pour chaque paire de voyelles synthétiques ainsi obtenues, les paramètres sont identiques: même fréquence de formant et bande passsante,même variation de F_0 et de l'énergie. L'effet de la source vocale sur le naturel du son de synthèse est étudié en utilisant des tests de perception. Les tests ont été effectués sur 25 personnes. Chaque auditeur est soumis à l'écoute de chacune des paires entre 10 et 30 fois dans un ordre aléatoire sous controle de l'ordinateur. Pour chacune d'elle, on pose la question: "quel est le son le plus

naturel?". Les résultats ont montré une nette préférence de la source élaborée pour toutes les voyelles (58 % à 72 %) à l'exception de la voyelle /u/ ou les pourcentages sont égaux. Par ailleurs, on constate que l'écart type des résultats varie de 20 % à 32 % (S. AWAD 1983). Ce résultat est normal dans ce type de test.

En conclusion, ces expériences confirment l'importance de la forme de l'onde d'excitation sur la qualité de la parole de synthèse. D'autre études sont en cours afin d'introduire les effets du couplage source- conduit vocal.

BIBLIOGRAPHIE

AL ANSARI (1981) Etude du fonctionnement et simulation temps réel d'un modèle de la source vocale.
Thèse de Docteur Ingénieur INP Grenoble.

S. AWAD (1983) Synthétiseur à formants en temps réel. Etude d'une source d'excitation élaborée. Codage optimun des paramètres de commande.
Thèse de Docteur Ingénieur INP Grenoble.

G. FANT (1979) Vocal source analysis. A progress report.
STL-QPSR 3/4 31-53

G. FANT (1980) Voice source dynamics.
STL-QPSR 2/3 17-37

SYNTHESE DE PLUSIEURS STYLES D'ELOCUTION: INVARIANCE ET VARIANTES PROSODIQUES

C. SORIN et K. BARTKOVA : CNET TSS/RCP - 22301 LANNION - France

---

Le premier but de notre étude est de faire varier le "style" de voix synthétique en ne modifiant que les paramètres suprasegmentaux c'est à dire, dans le cas de synthèse par diphones, en gardant le même dictionnaire de diphones.

Le premier jeu de règles suprasegmentales (décrit dans EMERARD 1977) était celui d'une locutrice non-professionnelle (style neutre) dont la voix avait également servi à la constitution du dictionnaire de diphones. La seule modification apportée a été une accélération systématique de 15% de la durée des trames de tous les diphones.

Les patrons prosodiques pour le deuxième jeu de règles ont été définis à partir de l'analyse d'une voix de locutrice professionnelle. Le corpus comportait 40 phrases de type publicitaire.

Les différences essentielles entre les deux types de patrons prosodiques sont les suivantes. En ce qui concerne l'intonation on observe une plus forte dynamique de variation de Fo au sein d'un même mot prosodique et une plus grande régularité dans la pente globale de décroissance de Fo à partir des points-cibles.

En ce qui concerne l'organisation rythmique, on relève un nombre plus important de pauses dues à l'insertion fréquente de pauses brèves entre l'adjectif et le nom et avant tout terme sémantiquement important.

Des règles de gestion des durées phonémiques ont également été introduites. Elles concernent essentiellement des allongements consonantiques en début de mots lexicaux, des allongements vocaliques en fin, des compressions des durées des groupes consonantiques et des voyelles au coeur des mots plurisyllabiques.

Une démonstration des deux styles de voix sera faite qui permettra de discuter des limites des variations prosodiques possibles pour des paramètres segmentaux figés.

Un deuxième but de cette étude est de disposer d'un module de règles prosodiques qui soit indépendant de la méthode de synthèse utilisée (diphones,formants). Or les règles de gestion de durée introduites dans les deux types de régles présentées ci-dessus sont des règles "ad hoc", dépendantes de la

méthode de synthèse utilisée (diphones). En effet, les diphones stockés présentent des variations intrinsèques de durée suivant qu'ils ont été extraits en début, fin ou milieu de mots. Ces variations, non contrôlées, induisent, après concaténation, des durées phonémiques hybrides.

Pour pallier cet inconvénient, une étude systématique des durées phonémiques en contexte, du français a été entreprise dans une optique "multi-style". L'étude des durées en cours porte sur 4 locuteurs choisis pour leurs vitesses d'articulation très différentes (de 4,2 à 6,1 syll/sec.). Deux corpus ont été utilisés pour chaque locuteur. Le premier corpus est constitué de logatomes monosyllabiques inclus dans des phrases porteuses: il comporte 174 logatomes de type "tVC" et 234 logatomes de type "aCV". Le deuxième corpus comprend 30 phrases, significatives, de 15 syllabes, prononcées sans pause.

Cette étude doit fournir une série de règles des gestion des durées phonémiques dont certaines sont indépendantes du "style de voix" recherché et d'autres spécifiques de débit lent ou plus rapide.

L'introduction de ces règles dans le système de synthèse reviendra à une regénération complète des durées phonémiques, en s'abstrayant des durées originelles de diphones stockés.

Une fois défini l'ensemble des patrons prosodiques utilisables pour un style de voix donnée,reste le problème de l'attribution automatique de ces patrons aux divers éléments du texte à synthétiser. Dans le cadre d'un système de dialogue homme-machine, ces règles pourront être incluses dans le module de génération automatique de messages, en utilisant l'information grammaticale, syntaxique et sémantique disponible à ce niveau.

UNE PAIRE DE PICS SPECTRAUX COMME CORRELAT ACOUSTIQUE DE LA NASALISATION DES VOYELLES

S. MAEDA : Centre National d'Etudes des Télécommunications,
22301 LANNION - France.

Une étude de simulation (exposée au Symposium Satellite de Toulouse, 11ème I.C.A., 1983) a montré que, lorsque le conduit nasal est couplé acoustiquement avec le conduit vocal principal par un abaissement du voile du palais, un pic spectral apparaît dans les basses fréquences pour les 11 voyelles testées. Le pic correspond au premier formant vocal (F1) combiné avec le formant glottique (FG) pour les voyelles fermées comme /i/ ou /u/, et au seul FG pour les voyelles ouvertes comme /a/. L'autre pic particulier aux voyelles nasalisées est soit, pour les voyelles postérieures comme /ɑ/ ou /u/, un formant (F2') qui est le deuxième formant (F2) augmenté en fréquence à cause du couplage, soit, pour les voyelles antérieures comme /i/, un formant nasal (FN). Nous nous intéressons, dans cette communication, aux deux types de pics centraux, un dans les très basses fréquences et l'autre dans les fréquences plus élevées.

Une telle paire semble le corrélat acoustique principal de la nasalisation des voyelles. Dans une expérience avec un synthétiseur à formants, par exemple, quand le F2 de la voyelle /u/ est décalé vers une fréquence plus haute, cette voyelle est perçue comme nasalisée. Quand les composants de hautes fréquences sont éliminées dans la voyelle /i/, synthétisée par simulation, le signal est interprété comme /u/, mais nasalisé. Au contraire, quand un des deux pics est éliminé par un filtrage approprié, passe-bas ou passe-haut, le signal n'est plus perçu comme nasalisé.

Une question importante est de savoir si une telle paire peut être produite seulement par couplage nasal. Si c'est le cas, la paire de pics doit être considérée comme l'indice qui signale la nasalisation, et probablement, le trait [NASAL]. Quand les paires de pics nasaux sont tracées sur le plan F1-F2, elles sont concentrées dans une zone étroite (une "zone nasale") entre /u/ et /y/. En effet,

les données sur les F1-F2 des voyelles pour beaucoup de langages indiquent un espace vide à cet endroit, et la zone nasale paraît bien dans cet espace. On tente d'assumer que cet espace vide est la conséquence directe des caractéristiques inhérentes de notre système du conduit vocal.

Afin d'évaluer cette hypothèse, nous avons fait une expérience avec un modèle articulatoire sur la base de données rayons X. Onze valeurs différentes pour chacun des cinq paramètres articulatoires (position de la mâchoire, du corps de la langue, et de l'apex de la langue, la forme dorsale, et les lèvres) sont espacées de façon linéaire. Les fréquences de F1 et F2 sont calculées pour toutes les combinaisons des valeurs des paramètres. Le résultat provisoire indique que la dispersion des points sur le plan F1-F2 forme un triangle, et que les points sont denses le long des zones correspondant à /i-e-a-o-u/ et moins denses à l'intérieur du triangle, y compris la "zone nasale". Si on se réfère à la nature quantique de la parole, proposée par Stevens, les voyelles doivent occuper les régions de forte densité, où les commandes articulatoires ne sont pas critiques pour la production d'une voyelle spécifique. Quoique les sons, avec une paire de pics spectraux situés dans la "zone nasale", pourraient être produits sans couplage nasal, ces sons sont prononcés par un moyen plus "approprié", c'est-à-dire, par l'abaissement du voile du palais, pour toutes les voyelles. En conclusion, donc, la paire de pics spectraux peut être un bon candidat pour le corrélat acoustique le plus important pour la nasalisation, et probablement pour le trait [NASAL].

EVALUATION PERCEPTIVE DE LA DETECTION DE VOISEMENT

A.C.M. RIETVELD, N.J.T. van ROSSUM: Instituut voor Fonetiek
Kath. Universiteit, NIJMEGEN - Pays-Bas.

L'un des problèmes que comporte l'évaluation de détecteurs de voisement est le manque d'un critère formel, permettant d'étiquetter un signal comme 'voisé' ou 'non-voisé'. Très souvent on se contente de comparer le résultat du détecteur au signal acoustique. Au milieu de voyelles cette procédure n'est pas difficile, mais, aux frontières de ces segments la situation est moins claire. C'est que le signal a tendance à perdre progressivement sa périodicité de sorte que les interprétations deviennent plus difficiles. L'évaluation d'un détecteur est d'autant plus difficile que les diverses méthodes d'évaluation possibles peuvent aboutir à des résultats différents. Sans aucun doute cette différence s'explique-t-elle par le fait qu'un rapport un sur un manque dans les domaines ou l'action des cordes vocales se manifeste.

Dans cette communication nous présenterons deux expériences sur l'évaluation des détecteurs de voisement. La première expérience consiste dans un 'scanning perceptif', l'autre dans la resynthèse de fragments de parole dans lesquels les transitions de voisement ont été déterminées à l'aïde de critères différents: les appréciations des auditeurs, qui ont été obtenues dans la première expérience, les décisions prises par un détecteur analogique, celles du détecteur digital SIFT (ILS) et celles de deux versions adaptées de ce dernier algorithme.

Dans la tâche de 'scanning perceptif' on a demandé aux auditeurs de catégoriser les segments de parole de 30 ms comme 'voisés' ou 'non-voisés'. Les appréciations obtenues dans cette tâche se sont avérées très fiables, bien que la confiance des jugements sur la fin des segments soit moins grande que celle des jugements sur le début (coefficient d'Ebel: .90 et .96 resp.).

Dans la seconde expérience, des versions différentes de parole resynthétisée ont été présentées à des auditeurs, qui ont été priés d'en apprécier la qualité relative; les versions ne différaient que dans la position des transitions de voisement. La qualité de la parole dans laquelle les segments voisés ont été déterminés par le détecteur de mélodie SIFT, mais raccourcis de 20 ms des deux côtés, a été considérée la plus acceptable. On a constaté que l'acceptabilité des versions dans lesquelles les transitions de voisement avaient été déterminées à l'aide du 'scanning perceptif' était à peine inférieure à celle du 'SIFT raccourci' (voir Fig 1a).

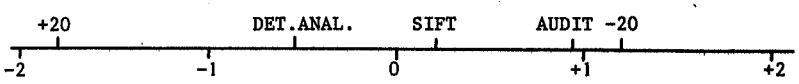


Fig 1a Echelle d'acceptabilité avec les positions des cinq versions de parole resynthétisée

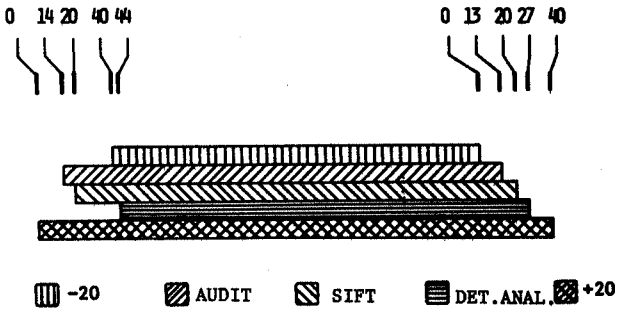


Fig 1b Les moyennes des transitions non-voisé/voisé et voisé/non-voisé en millisecondes avec point de référence arbitraire

En comparant la Fig 1a avec Fig 1b, on constate une corrélation absolue entre les positions relatives des transitions voisé/non-voisé et les positions des différentes versions sur l'échelle d'acceptabilité: l'allongement final des segments voisés a un effet nettement négatif sur l'acceptabilité des versions en question, alors que l'effet de la position des transitions non-voisé/voisé est moins clair.

Des expériences sont en cours afin de déterminer les effets relatifs de raccourcissement et d'allongement initial ou final de segments voisés sur la qualité de la parole resynthétisée.

PROSODIE ET INTEGRATION PHONETIQUE DU SILENCE

N. BACRI: Centre d'Etude des processus cognitifs et du langage, CNRS-EHESS, PARIS
A. NICAISE: Dép. de Rech. Ling., Univ. Paris VII et Univ. Paris XII, PARIS.

Un silence peut être un indice pour la perception des occlusives et de la différence fricatives / affriquées (Dorman et al., 1980), mais une affriquée n'est pas perçue si le silence coïncide avec un changement de locuteur ou une limite de phrase (Rakerd et al., 1982). La continuité articulatoire expliquerait ces résultats. Elle correspond à un faisceau d'indices acoustiques: durée vocalique, configuration de F_o, mouvement du deuxième formant, présence et durée d'un silence, offset de la voyelle, attaque et durée de la fricative suivante. Certains de ces indices peuvent recevoir 2 interprétations en fonction de leur valeur. Ainsi le silence au niveau phonétique est interprété comme la tenue d'une occlusive ou comme une pause. Par ailleurs, certaines limites syntaxiques (principale / subordonnée en si) sont marquées par des contours intonatifs différents: continuation mineure (Cm) ou majeure (CM), contour final suivi d'une parenthétique (Pa).

On fait l'hypothèse que certains de ces contours s'accompagnent d'une rupture de la continuité articulatoire (CM et Pa), d'autres non (Cm), et que les modalités d'intégration perceptuelle d'un silence post-vocalique varieront en fonction de la configuration de F_o.

On a tout d'abord établi l'existence de 2 pôles de l'intégration phonétique du silence en français. Les sujets avaient à identifier un énoncé resynthétisé par LPC, comportant différentes durées de silence (0 à 70 ms.par pas de 15 ms.;100,150,200 ms.). Le silence suivait la voyelle du mot fait, susceptible d'être interprété comme fête si l'occlusive est perçue, et précédait la fricative [s] (si). L'intégration du silence dans la perception d'une occlusive apparaît entre 20 et 45 ms., en Cm.

Quand on part d'un contour CM (voyelle finale allongée, montée F_o de 175 à 240 Hz, silence de 250 ms.), la perception d'une occlusive ne peut être obtenue par réduction de la pause à 50 ms. et abrégement de la voyelle (cf. Rakerd et al.).

Dans la deuxième expérience, il s'agit d'analyser si certains des paramètres acoustiques associés à la perception d'une rupture de la continuité articulatoire peuvent à eux seuls déterminer cette perception de rupture (le silence serait perçu comme une pause) (Price et Levitt, 1983). Partant d'un énoncé du type précèdent, avec une Cm, on a construit 2 autres séries de stimuli par modification des valeurs de F_o comme des approximations des contours CM et Pa. La durée du silence a varié de 0 à 100 ms. par pas de 15 ms., de 100 à 220 ms. par pas de 30 ms. L'ordre de

présentation (procédure A B X, A et B distants de 30 ms.) était aléatoire.

La discrimination a été bonne (80% de réponses correctes) pour les contrastes entre silences inférieurs à 40 ms. en CM et Pa, 55 ms. en Cm, l'un seulement des stimuli A ou B entrainant la perception d'un [t] post-vocalique. Pour les valeurs du silence $\geqslant$70 ms., le contour Cm induit un 2$^{\text{ème}}$ pic de la fonction de discrimination entre 70 et 130 ms., pic situé aux alentours de 190 ms. en CM, les réponses en Pa demeurant aléatoires.

Une rupture de la continuité articulatoire, construite sur les contours F_o, n'entraine donc pas la suppression de l'intégration phonètique du silence, mais un déplacement des valeurs de la fonction de discrimination. Le 2$^{\text{ème}}$ pic suggère que la discrimination est facilitée quand le contraste entre les durées du silence est perçu comme l'opposition entre un [t] post-vocalique [fɛt] et un [t] pré-fricatif [tsi], non plausible dans le contexte de l'énoncé étudié. Cette opposition se produit pour des valeurs de silence d'autant moins longues que le silence ne peut être interprété comme une pause en raison de la présence d'indices en compétition, en Cm (F_o de continuation, long silence). Par contre un silence est perçu comme une tenue du [t] post-vocalique jusqu'à 160 ms. lorsque les indices de rupture concordent (CM, Pa).

Les éventuelles relations d'échange entre contours F_o, durée du silence et durée vocalique sont en cours d'étude.

LE DEVELOPPEMENT DE L'HABILETE D'ANALYSE PHONETIQUE EXPLICITE DE LA PAROLE

A. CONTENT et J. MORAIS

Laboratoire de Psychologie Expérimentale - Université Libre de Bruxelles.

Comment l'enfant pré-lettré se représente-t-il les sons de la parole ? Est-il, en particulier, capable de concevoir une expression comme une séquence de segments phonétiques ? Outre leur intérêt théorique, ces questions ont une importance pratique pour l'apprentissage de la lecture, puisque le système alphabétique est une représentation phonologique de la langue. Plusieurs études développementales basées sur des paradigmes différents ont montré des performances très faibles chez l'enfant de moins de six ans. De plus, comparant des adultes analphabètes et des adultes de même provenance qui avaient appris à lire tardivement, à l'aide d'une tâche qui consistait à répéter des expressions après en avoir supprimé la consonne initiale, Morais et al (1979) ont trouvé que les illettrés avaient des performances très faibles, similaires à celles d'enfants prélettrés, alors que les lecteurs réussissaient facilement.Une interprétation de ces résultats est que la représentation de la parole en phones est un produit de la connaissance du système alphabétique. Une autre hypothèse, préférée par les auteurs, est que la conscience des unités phonétiques n'apparaît qu'à la condition d'être confronté à une situation où elle est nécessaire. Les illettrés,adultes ou enfants, seraient donc en principe capables de découvrir les unités phonétiques indépendamment de l'acquisition de la lecture.

Nous avons exploré cette possibilité à l'aide de la même tâche, auprès d'enfants de quatre à six ans qui n'avaient aucune connaissance du principe alphabétique. Dans une première expérience, nous avons pu montrer que les enfants de cinq ans progressaient sensiblement en cours d'épreuve, lorsque de l'information corrective leur était fournie à chaque essai. Par contre, dans un groupe comparable de sujets d'un an plus jeunes, aucun progrès n'était observé. Dans une seconde expérience, décrite plus explicitement, nous avons comparé la suppression de la consonnes initiales et de consonnes finales. A chaque niveau d'âge (4 et 5 ans), deux groupes appariés ont été constitués sur base de la performance pour une épreuve de substitution de voyelles (ex. PAK-POK). Chaque groupe passait ensuite deux épreuves portant soit sur les consonnes initiales, soit sur les consonnes finales. Dans une des épreuves, il fallait produire la partie des items restant après suppression de la consonne critique, et dans l'autre, il fallait choisir parmi deux items celui qui commençait (ou finissait)

par la même consonne que le stimulus. Des progrès sensibles ont été observés au cours de l'épreuve de suppression, mais contrairement aux résultats précédents, aucune différence développementale n'a été observée. La suppression de la consonne initiale est apparue plus difficile que celle de la consonne finale, et plus affectée par la structure segmentale des stimuli.Cette différence ne s'explique pas par une plus forte saillance perceptive des consonnes finales. En effet, dans la tâche de choix binaire, c'est la classification sur base de la consonne initiale qui s'est avérée plus facile. L'interprétation proposée est que dans les deux tâches, une partie initiale pourrait être isolée plus facilement qu'une partie finale.

Pris ensemble, ces résultats suggèrent que les enfants pré-lettrés peuvent découvrir les segments phonétiques de la parole, et que cette prise de conscience ne nécessite pas la connaissance du code alphabétique.

Référence

Morais,J., Cary, L., Alegria, J. et Bertelson, P. (1979) Does awareness of speech as a sequence of phones arise spontaneously ? Cognition, 8, 323-331.

CODAGE DE L'INFORMATION VERBALE CHEZ LE DEFICIENT AUDITIF

J. LEYBAERT et J. ALEGRIA

Laboratoire de Psychologie Expérimentale - Université Libre de Bruxelles.

Outre la fonction communicative évidente que possède la parole, celle-ci joue un rôle extrêmement important au niveau de la communication de l'individu avec lui-même. Une partie des activités mentales des sujets, planifier, faire des calculs mentalement, etc., utilise la parole interne comme code de base. Dans le processus de lecture, les codes phonologiques jouent également un rôle important. Une distinction désormais bien établie est effectuée entre le code phonologique pré-lexical, qui permet d'accéder aux informations en mémoire dans la compréhension de la parole aussi bien qu'en lecture, et le code phonologique post-lexical, qui se produirait après qu'une entrée lexicale ait été atteinte. Des arguments empiriques mettant en évidence l'existence de ces deux formes de parole interne ont été recueillis en particulier au niveau de la lecture (chez l'adulte et l'enfant) et l'importance relative de ces deux processus est une des questions les plus discutées actuellement en psychologie cognitive (1).

On peut s'interroger sur la nature des codes qu'utilisent dans leurs activités mentales des sourds profonds de naissance ayant des compétences verbales limitées, ainsi que sur le rapport existant entre la qualité de leur parole externe et la présence ou non d'une parole interne identifiable.

Chez l'enfant entendant, l'utilisation du code phonologique pré-lexical est intéressante parce que celui-ci permet au sujet de se relier à une voie d'accès au lexique déjà abondamment utilisée au cours de son développement linguistique oral. On peut par contre penser que cette voie d'accès au lexique n'est pas disponible chez l'enfant sourd ayant une expérience auditive extrêmement appauvrie. Nous avons obtenu des résultats supportant cette hypothèse, au moyen d'une tâche de décision lexicale. L'expérience consistait à présenter aux sujets une série de lettres et à leur demander de décider le plus rapidement possible s'il s'agit d'un mot ou non. On constate chez les entendants que la réponse "non" à des pseudo-mots homophones de mots réels (par exemple GRYPHE) est plus lente que dans une condition contrôle. Ce phénomène est absent chez les sourds, même très compétents sur le plan oral.

On peut penser que l'entraînement intensif à la parole externe dont bénéficient les sourds éduqués dans une optique oraliste doit avoir pour conséquence, au moins pour certains d'entre eux, l'intériorisation de la prononciation des

mots, c'est-à-dire l'association de cette caractéristique aux autres attributs stockés dans l'entrée lexicale associée à un mot. Ceci leur permettrait de manipuler ces représentations phonologiques post-lexicales pour accomplir certaines activités cognitives. Classiquement, ce problème a été étudié par le biais d'expériences de mémorisation à court terme. On présente aux sujets une série d'items (lettres, chiffres, mots, dessins, etc.) qu'il doit conserver en mémoire pendant un temps bref. Afin d'éviter la verbalisation forcée, la présentation des stimulis est visuelle et la réponse écrite. L'analyse des erreurs montre que l'entendant utilise un code mnémonique verbal : les erreurs sont similaires de la réponse correcte au niveau phonologique. L'utilisation d'un tel code par les sourds est fortement corrélée avec une bonne intelligibilité de leur parole (2). Nous avons obtenu des résultats convergents à l'aide d'un autre paradigme expérimental, qui teste l'activation automatique du nom du mot une fois que celui-ci a été reconnu. Nos résultats montrent que ce processus est surtout caractéristique des sourds possédant une parole intelligible (3). Les résultats d'une nouvelle expérience, récemment terminée, devraient nous permettre de préciser le lien entre ces ces deux formes de codage phonologique post-lexical, à savoir l'activation automatique de la représentation phonologique associée au mot et la manipulation de ces représentations pour accomplir une tâche de mémorisation à court terme.

En résumé, il semble que l'utilisation de codes phonologiques dans leurs activités mentales soit limitée chez les sourds profonds de naissance, et ce à deux niveaux au moins. D'une part, les sourds qui arrivent à acquérir une parole externe de bonne qualité sont les seuls à utiliser un code de ce type. D'autre part, un tel code ne participe pas chez ces sujets au processus d'accès lexical, mais n'est disponible qu'une fois que celui-ci a été réalisé.

Références

(1) McCusker, L.X., Hillinger, M.L., & Bias, R.G. Phonological recoding and reading. Psychological Bulletin, 1981, 89, 217-245.

Jorm, A.F. & Share, D.L. Phonological recoding and reading acquisition. Applied Psycholinguistics, 1983, 4, 103-147.

(2) Conrad, R. The deaf school child. London : Harper & Row, 1979.

(3) Leybaert, J., Alegria, J. & Fonck, E. Automaticity in word recognition and word naming by the deaf. Cahiers de Psychologie Cognitive, 1983, 3, 255-272.

LISTE PROVISOIRE DES PARTICIPANTS

ABRY C.	I.C.P. GRENOBLE
ADDA G.	LIMSI-CNRS ORSAY
AGGOUN A.	CNET LANNION
ALEGRIA J.	U.L.B. BRUXELLES
ALIF S.	L.C.T. VILLACOUBLAY
ALINAT P.	THOMSON-CSF. CAGNES-SUR-MER
ANDREEWSKY A.	LIMSI-CNRS ORSAY
AWAD S.	I.C.P. GRENOBLE
BACRI N.	CNRS PARIS
BARTKOVA K.	CNET LANNION
BEECKMANS R.	U.L.B. BRUXELLES
BENOIT C.	I.C.P. GRENOBLE
BEROULE D.	LIMSI-CNRS ORSAY
BERTELSON P.	U.L.B. BRUXELLES
BIETRY C.	LIMSI-CNRS ORSAY
BOE JL.	I.C.P. GRENOBLE
BONNOT JF.	UNIV. AMSTERDAM HOORN
BOULOGNE M.	I.C.P. GRENOBLE
BREANT M.	UNIV. PAUL SABATIER TOULOUSE
CAELEN J.	UNIV. PAUL SABATIER TOULOUSE
CAELEN-HAUMONT	UNIV. PAUL SABATIER TOULOUSE
CARBONELL N.	CRIN VANDOEUVRE
CARRE R.	E.N.S.E.R.G. GRENOBLE
CARRIER A.	LIMSI-CNRS ORSAY
CARTIER M.	CNET LANNION
CARTON F.	UNIV. NANCY MALZEVILLE
CAVE C.	UNIV. PROVENCE AIX EN PROVENCE
CHAFCOULOFF M.	INST. PHON. AIX EN PROVENCE
CHARPENTIER F.	CNET LANNION
CHAZE JCH.	CEA GIF/YVETTE
CHEVALLIER MC.	ENST PARIS
CHEVRIE C.	I.N.S.E.R.M. U3 PARIS
CHOLLET G.	E.N.S.T. PARIS
CONTENT A.	U.L.B. BRUXELLES
COURBON JL .	CNET LANNION
DABOUZ M.	E.N.S.T. PARIS
DAMESTOY JP.	CRIN NANCY

DE BOYSSON B.	LAB. PSYCHO.MSH PARIS
DE FRANCE R.	LIMSI-CNRS ORSAY
DEBOULE D.	LIMSI-CNRS ORSAY
DELGUTTE B.	CNET LANNION
DEROUAULT AM.	IBM. PARIS
DESCOUT R.	CNET LANNION
DESI M.	CHU LE KREMLIN-BICETRE
DEVOOGHT J.	U.L.B. BRUXELLES
DI CRISTO A.	INST. PHON. AIX EN PROVENCE
DIERICKX J.	U.L.B. BRUXELLES
DISSOUBRAY R.	ANTIBES
DOLMAZON JM.	INST. COM. PARLEE GRENOBLE
DOLOGLOU Y.	I.C.P. GRENOBLE
DONHOUEDE B.	E.N.S.E.R.G. GRENOBLE
DUBOIS D.	CNET LANNION
DUMOUCHEL P.	I.N.R.S. MONTREAL
DUPEYRAT B.	CEA GIF/YVETTE
ESCUDIER P.	E.N.S.E.R.G. GRENOBLE
ESKENAZI M.	LIMSI-CNRS ORSAY
ESPESSER R.	INST. PHON. AIX EN PROVENCE
EVRARD D.	CNET LANNION
FENG G.	I.C.P. GRENOBLE
FLOCON B.	C.G.E. MARCOUSSIS
FLUHR C.	UNIV. PARIS-SUD ORSAY
FOHR D.	CRIN NANCY
FRAILONG JM.	E.N.S. PARIS
FRAYSSE B.	UNIV. PAUL SABATIER TOULOUSE
GAGNOULET C.	CNET LANNION
GALLAIS A.	THOMSON-CSF CAGNES-SUR-MER
GANGOLF JJ.	LIMSI-CNRS ORSAY
GAUVAIN JL.	LIMSI-CNRS ORSAY
GAY T.	UNIV. CONN. FARMINGTON
GENTIL M.	IBM PARIS
GIBIAT V.	UNIV. PARIS II ORSAY
GILLOUX M.	CNET LANNION
GISPERT J.	FAC. SC. LUMINY MARSEILLE
GOUDAILLIER JP.	UNIV. R.DESCARTES PARIS
GREINER G.	I.N.S.E.R.M. U3 PARIS
GRENEZ F.	U.L.B. BRUXELLES
GRENIER Y.	E.N.S.T. PARIS
GROSSETETE D.	C.G.E. MARCOUSSIS
GUERIN B.	E.N.S.E.R.G GRENOBLE
GUIDET C.	I.N.S.E.R.M. U3 PARIS
GUIZOL J.	FAC. SC. LUMINY MARSEILLE
HALLE P.	LIMSI-CNRS ORSAY
HARMEGNIES B.	UNIV. MONS
HATON JP.	UNIV. C.R.I. NANCY
HESS W.	TECHN. UNIV. MUNICH
HOMBERT JM.	UNIV. LYON2 LYON
INDEFREY H.	TECHN. UNIV. MUNICH
JANDOT J.	UNIV. PARIS VII FRONTIGNAN
JARDIN P.	UNIV. PARIS II ORSAY

JOSPA P.	U.L.B. BRUXELLES
JOUVET D.	CNET LANNION
KONOPCZYNSKI G.	INST. PHON. STRASBOURG
KOSTER JP.	UNIV. TREVES
LANDERCY A.	UNIV. MONS
LEFEVRE JP.	CIT LA VERRIERE
LEYBAERT J.	U.L.B. BRUXELLES
LHERM B.	I.C.P. GRENOBLE
LHOTE E.	UNIV. BESANCON
LIENARD JS.	LIMSI-CNRS ORSAY
LOCKWOOD P.	C.G.E. MARCOUSSIS
LONCHAMP F.	UNIV. NANCY
MAEDA S.	CNET LANNION
MAJANI E.	LIMSI-CNRS ORSAY
MARIANI J.	LIMSI-CNRS ORSAY
MARTIN P.	UNIV. TORONTO
MATON B.	I.N.S.E.R.M. U3 PARIS
MELONI H.	FAC. SC. LUMINY MARSEILLE
MERCIER G.	CNET LANNION
MERLIER B.	I.C.P. GRENOBLE
MICLET L.	E.N.S.T. PARIS
MORAIS J.	U.L.B. BRUXELLES
MOREL C.	C.C.E.T.T. RENNES
MOURADI A.	LABO ELEC. RABAT
MOUSEL P.	CRIN NANCY
NAJIM M.	LABO ELEC. RABAT
NEEL F.	LIMSI-CNRS ORSAY
NICAISE A.	DRL GIF/YVETTE
PASCAL D.	CNET LANNION
PASSIEN O.	CIT LA VERRIERE
PIERREL JM.	CRIN NANCY
PISTER C.	CRIN NANCY
POIRIER F.	LIMSI-CNRS ORSAY
PRICE P.	CNET LANNION
PUECH G.	UNIV. LYON2 LYON
QUENOT G.	LIMSI-CNRS ORSAY
RAJOUANI A.	LABO ELEC. RABAT
REICHART A.	LIMSI-CNRS ORSAY
RIETVELD T.	INST. PHON. NIJMEGEN
SAGARD L.	LIMSI-CNRS ORSAY
SANCHEZ H.	I.C.P. GRENOBLE
SANTERRE L.	UNIV. MONTREAL
SAP J.	C.G.E. MARCOUSSIS
SAUTER L.	C.G.E. MARCOUSSIS
SCHOENTGEN J.	U.L.B. BRUXELLES
SCHWARTZ JL.	E.N.S.E.R.G. GRENOBLE
SERNICLAES W.	U.L.B. BRUXELLES
SIMON MA.	CNET LANNION
SIMON P.	INST. PHON. STRASBOURG
SORIN C.	CNET LANNION
SOUDOPLATOFF S.	IBM PARIS
SOURY JL.	LIMSI-CNRS ORSAY

STELLA M.	CNET LANNION
SYLIN G.	U.L.B. BRUXELLES
TARRIDEC C.	CNET LANNION
TESTON B.	INST. PHON. MARSEILLE
THOUZERY J.	SILEC PARIS
TUAL D.	CEA-DEIN GIF/YVETTE
TUBACH JP.	E.N.S.T. PARIS
URGELL H.	UNIV. PAUL SABATIER TOULOUSE
VAISSIERE J.	CNET LANNION
VAN ROSSUM NJT.	KATH. UNIV. NIJMEGEN
VIARA E.	CIT LA VERRIERE
VILACLARA G.	EC. POLYTECH. LAUSANNE
WAJSKOP M.	U.L.B. BRUXELLES
WU F.	UNIV. PARIS II ORSAY
ZANELLATO G.	UNIV. MONS
ZERLING JP.	UNIV. STRASBOURG

Imprimé par les Presses Universitaires de Bruxelles, a.s.b.l.